DOCUMENT RESUME

ED 414 997 PS 025 428

AUTHOR Kagan, Sharon Lynn, Ed.; Rosenkoetter, Sharon, Ed.; Cohen, Nancy, Ed.

TITLE Considering Child-Based Results for Young Children: Definitions, Desirability, Feasibility, and Next Steps.

INSTITUTION Yale Univ., New Haven, CT. Bush Center in Child Development and Social Policy.

PUB DATE 1997-00-00

NOTE 81p.; Based on Issues Forums on Child-Based Outcomes (New York, NY, June 1-2, 1995 and January 24, 1996).

PUB TYPE Information Analyses (070) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC04 Plus Postage.

DESCRIPTORS *Academic Standards; Educational Change; *Educational Improvement; *Educational Objectives; Feasibility Studies; *Outcome Based Education; *Outcomes of Education; Preschool Education; Program Effectiveness; Theory Practice Relationship

ABSTRACT

While elementary and secondary educators are pursuing a results orientation as a means of educational reform and pedagogical improvement, the early childhood field has not explored sufficiently the desirability and feasibility of, nor the process for, establishing child-based standards and results for younger children. This report presents a synthesis of issues discussed at two issues forums held in 1995 and 1996. Following an introductory chapter, chapter 2 distinguishes between types of results and purposes of results as they apply to young children. Chapter 3 examines the desirability of child-based results in early childhood education in terms of the impact on teachers' practice and children's experiences, public understanding, funding, and the relationship of early care and education to other services. Chapter 4 identifies the following five conditions necessary for child-based results to be feasible: broad participation in the identification of results; identification of appropriate results; clarity concerning which children to include; appropriate measurement of results; and linking of child-based results to efforts to improve the lives of children. Chapter 5 outlines steps to advance a results-based approach: increase public consciousness and participation; plan strategically; identify and choose results carefully; develop appropriate, cost-effective approaches to assessment and data collection; put theory into practice; explore ways to adequately fund a results approach; and communicate, implement, and evaluate such programs. Five appendices contain the meeting agendas of each of the issues forums and three papers presented at the meetings. (TJQ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 414 997

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Considering CHILD-BASED RESULTS for Young Children

Definitions, Desirability, Feasibility, & Next Steps

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
S.L. Kagan
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

PS 025428

Based on Issues Forums on Child-Based Outcomes sponsored by The W.K. Kellogg Foundation, The Carnegie Corporation of New York, and *Quality 2000:* Advancing Early Care and Education

BEST COPY AVAILABLE

Considering Child-Based Results for Young Children

Definitions, Desirability, Feasibility, and Next Steps

Based on Issues Forums on Child-Based Results
sponsored by
The W. K. Kellogg Foundation,
The Carnegie Corporation of New York, and
Quality 2000: Advancing Early Care and Education

Edited by
Sharon Lynn Kagan,
Sharon Rosenkoetter, and
Nancy Cohen

©1997
Yale Bush Center in Child Development and Social Policy
310 Prospect Street
New Haven, CT 06511

Contents

1. INTRODUCTION

Background 1
Rationale for and Goals of the Issues Forums 3

2. DEFINITIONS

Defining Different Types of Results 4
Defining Different Purposes for Assessing Results 6
Issues Concerning Definitions of Types and Purposes of Results 7

3. DESIRABILITY

The Impact on Teachers' Practice and Children's Experiences 10
The Impact on Public Understanding 10
The Impact on Funding 11
The Impact on the Relationship of Early Care and Education and Other Services 12
So...What do we Conclude About the Desirability of Moving to a Child-Based Results Approach? 12

4. FEASIBILITY

Only if There is Broad Participation in the Identification of Results 14
Only if we can Identify Appropriate Results 14
Only if we are Clear About Which Children to Include 16
Only if we Measure Results Appropriately 17
Only if we Link Child-Based Results to Efforts to Improve the Lives of Children 17

5. SUGGESTED NEXT STEPS

Increase Public Consciousness and Participation 19
Plan Strategically 19
Identify and Choose Results Carefully 20
Develop Appropriate, Cost-Effective Approaches to Assessment and Data Collection 21
Put Theory Into Practice 21
Explore Ways to Fund a Results Approach Adequately 22
Communicate, Implement, and Evaluate 22

6. REFERENCES

7. APPENDICES

A. June 1–2, 1995 Meeting Agenda and Participants 24
B. January 24, 1996 Meeting Agenda and Participants 27
C. Considerations Regarding an Outcomes-Based Orientation, Sheldon H. White 29
D. Judging Interventions by Their Results, Lisbeth B. Schorr 36
E. Can we Measure the Outcomes?, John M. Love 48

Chapter 1. *Introduction*

While elementary and secondary educators are pursuing results as a means of educational reform and pedagogical improvement, the early childhood field has not explored sufficiently the desirability, feasibility, or process of establishing child-based standards and results for younger children. This report presents a synthesis of issues discussed at two Issues Forums. The first, on Child-Based Results, was held on June 1–2, 1995 in New York City and was attended by 37 scholars and practitioners in early care and education, school reform, and policy development related to children and families (see Appendix A). The second forum, held on January 24, 1996, addressed Next Steps in Advancing Child-Based Results; it consisted of 19 participants with expertise in results efforts (see Appendix B). The Forums were a collaborative effort of the W.K. Kellogg Foundation, Carnegie Corporation of New York, and *Quality 2000:* Advancing Early Care and Education.

Background

Current challenges facing the development and implementation of child-based results for young children are framed by two factors, each discussed below. The first is the current socio-political context; the second derives from grave concerns regarding the challenges and misuse of results in the past.

THE PRESENT CONTEXT

Increasing dissatisfaction with America's schools and the performance of its graduates has fostered widespread calls for educational reform. Emanating from both the public and private sectors, dissatisfaction with American education is compounded by growing concerns about the rising costs of a system perceived to be inefficient as well as ineffective.

To begin to rectify these educational ills, several movements have taken hold, including school-based management, charter schools, standards specification, and results-driven accountability. The results movement gained considerable public attention through the Goals 2000: Educate America Act and the Elementary and Secondary Education Act, among others. The press for greater accountability has spurred federal and state action, with new standards and results groups being formed to address educational challenges. In some states—Vermont, Minnesota, and Oregon—the results orientation transcends education and extends across the array of human services.

As support for a results orientation in education and human services increases, so has attention to young children and their families. New policy initiatives focus on very young children (e.g., Early Head Start and Healthy Start), preschoolers (e.g., National Governors' Association Action Teams on School Readiness), and on both (e.g., Kids Count, Child Care and Development Fund, and the family support movement). Public schools increasingly support prekindergarten services, while programs such as Head Start and Parents as Teachers receive media consideration and greater funding.

Despite dramatic growth in early care and education and unprecedented calls by the National Education Goals Panel and others to delineate optimal results and chronicle children's progress toward them throughout the nation, the field of early care and education has responded with less than vigorous support. Indeed, calls for a focus on child-based results have met with staunch vocal resistance, as well as more silent pleas for a redirection of effort.

THE PAST CONTEXT

Reluctance to embrace a results orientation by the early care and education field has deep roots, including legitimate concerns regarding test misuse, technical concerns regarding measurement, a historic focus on process, and a lack of agreement regarding what is meant by results.

Evidence of Misuse of Test Data

Early childhood education professionals have criticized the use of tests to mis-label, mis-categorize, and stigmatize children during their earliest days in formal education (Meisels, 1988; Shepard & Smith, 1987), and they have questioned the validity of standardized tests for individual children—especially boys, racial minorities, and preschool children whose primary language is not English. Of great concern to the field of early care and education has been the widespread use of "readiness" instruments to screen children for entry to school. This practice has resulted in up to 50 percent of children in some districts delaying school entry or being sent to alternative "transition" classes of unsubstantiated value (Gnezda & Bolig, 1989; Graue, 1993). Given these experiences, there is well-grounded skepticism in the early childhood education community about the potential use and misuse of results.

Concerns About Measurement

Concerns about measurement manifest themselves in two domains; what is measured and how it is measured. In a paper commissioned for the first Forum, White considers some of these concerns (see Appendix C). Regarding what is measured, early educators worry that child results may be narrowly constructed to include only cognitive and pre-academic results, ignoring developmental domains that are crucial to children's success but more difficult to capture in routinized assessment (e.g., socio-emotional development and approaches toward learning). Regarding the measurement of results—the how—some early educators and developmentalists doubt the feasibility of instituting a results approach with young children due to the variability of their behavior and their inexperience in "performing" in testing situations. Because young children's learning is highly episodic, early educators voice concerns regarding the capacity of instruments administered to children on one occasion to capture developmental nuances accurately. Further, they worry that assessments may not give racially, ethnically, and linguistically diverse young children appropriate opportunities to display their skills and knowledge. Finally, they challenge the reliability and validity of existing assessment tools, questioning whether such instruments can be suitably altered or new instruments created, to diminish these concerns.

Focus on Process

Early education historically has emphasized the process of young children's individual learning. Early educators are trained to recognize and work with children's uneven growth as well as the diversity in family values, experiences, and interaction styles that shape early development. Reflecting this individualistic orientation, historic attempts to codify and improve practice in the early childhood education field have focused on the modification of inputs, structural variables, and process variables that enable such individualization—adult-child ratios, group size, and interaction patterns. Routinely, such inputs have been equated with quality, with little call for an examination of the need for a results orientation. There is little press for a movement toward child-based accountability by the early childhood education field, particularly when child-based results could be used to influence critical program funding and policy decisions.

Lack of Clarity of Terms

Finally, discourse on child results to date has been hampered by imprecision in language and scope.

There is limited consensus in early care and education regarding what is meant by terms in common use—goals, benchmarks, results, inputs, indicators, interim indicators, assessment, and testing. Attempts to achieve definitional clarity have been overshadowed by the field's dual concern with the delivery of direct services to children and their families, on the one hand, and with building the infrastructure on the other hand. In short, definitional ambiguity is pervasive.

Rationale for and Goals of the Issues Forums

Given this context, and the likelihood of ongoing public pressure for results for young children, it was deemed appropriate to engage scholars, practitioners, and policymakers in a professional conversation. Organizers of the Forums wished to provide an opportunity to take stock of the current status of child-based results for children birth to age eight, and to give voice to the early childhood education community regarding its issues and concerns.

In particular, it seemed important to: (1) clarify definitional distinctions; (2) discern the desirability of moving to a results orientation; (3) determine the feasibility of moving toward a results orientation for young children; (4) consider next steps regarding a results-based orientation. In contrast, the aim of the Forums was not to consider or define the specific content of results that might be deemed appropriate for young children; that work has been started by others (Love, Aber, & Brooks-Gunn, 1994; Phillips & Love, 1994; Institute for Research on Poverty, 1995). Nor, given the complexity of the issues and the differences of opinion that exist, was it the intent of the Forums to achieve consensus on relevant issues; rather, this was an opportunity for honest reflection and thoughtful debate.

Reflecting these goals and intents, this report is structured around three themes—definitions, desirability, and feasibility—with possible next steps suggested by the participants at the end. The document represents a synthesis of the issues discussed as well as those documented in the literature. It is not intended to represent the consensus of participants—because no such consensus was achieved. Rather, it is intended to reflect the complexity of the issues and the challenges associated with moving toward child-based results for young children. Editors of the document (Sharon L. Kagan, Sharon Rosenkoetter, and Nancy Cohen) have attempted to represent the ideas with fidelity; they alone, however, are responsible for errors. Appendices follow, including meeting agendas, lists of participants, and working papers prepared for the first Forum by Sheldon White, Lisbeth Schorr, and John Love.

Chapter 2. *Definitions*

This section details two distinct, but related, frameworks that shape the discussion that follows. The first distinguishes among different *types* of results; the second distinguishes among different *purposes* of results. A third part of this section discusses issues related to both the types and purposes of results. It should be noted that the definitions offered are not the only way of distinguishing among the types or purposes of results (Bruner, Bell, Brindis, Chang, & Scarbrough, 1993; Young, Gardner & Coley, 1993; Schorr, 1994); they simply represent one heuristic. Important to note, however, is the pervasive lack of consensus around definitions as well as the need to ground this (and other) discussions of results in a definitional framework. **It should be noted that while specification of various types and purposes of results is critical for clear communication in this discussion, the deliberations of the Forums were designed to focus on *Type One Results* (what children know and can do and what is hereafter referred to as "child-based results") and on *Purpose Four* (accountability).**

Defining Different Types of Results

Four types of results have been identified. Each type is discernable and knowable, each demands its own data elements and approaches to collection, and each evokes its own assessment processes and considerations. Each type, though independent and distinct, can be used in concert with others, contingent upon the purposes of the data collection. Together, the types of results form a continuum, with items representing children's performance and behavior at one end, and results related to systemic performance at the other end. In concert, the four types represent a comprehensive overview of the kinds of information being considered by agencies, localities, and states as they move to a results orientation (Kagan, 1995). Data on what children know and can do can be considered primary results, while secondary results include the contexts in which children develop, such as child and family conditions, service provision and access, and systems capacity.

TYPE ONE RESULTS—WHAT CHILDREN KNOW AND CAN DO

This type of information focuses directly on children's *behaviors*—what children know and can do. It is synonymous with the term "child-based results," the focus of this document. This type of information must be gathered by observing children directly. It accepts no proxies for behavior, but is a precise and accurate description of children's performance. For young children this includes dimensions related to their motor development, social and emotional development, use of language, cognition and general knowledge, and approaches to learning. To gather this type of information, child behavior is typically recorded intermittently, from more than one data source.

Examples of Type One Results include:

Motor development: Prevalence of children: who jump; walk a six-foot balance beam; cut; do "x" piece puzzle.

Social and emotional development: Prevalence of children: who accept responsibility for own actions; take turns; form and maintain friendships.

Language usage: Prevalence of children: who initiate and sustain conversation; listen to others; recite poems and do fingerplays; repeat a sentence in correct word order; follow verbal direction containing three steps; tell about a picture when looking at it; name common objects.

Cognition and general knowledge: Prevalence of children: who match; sort shapes and colors; identify largest and smallest; demonstrate awareness of cause and effect.

Approaches toward learning: Prevalence of children: who take risks; persevere in a chosen activity; demonstrate curiosity; use materials in inventive ways.

TYPE TWO RESULTS—CHILD AND FAMILY CONDITIONS

This type of results focuses on the *conditions* that surround and encase what children know and can do. Such information may be gathered from reviews of documents, including health records; interviews with family members and service providers; and direct observations/conversations with children and their families. This type of results assumes that what children know and can do is directly related to their own health status and to the conditions in which they live. Rather than reporting data on individual children, this type of data is generally reported as aggregated prevalences and percentages. Child and family results may be grouped into categories (e.g., child health conditions; family income conditions) with positive and negative indicators in each.

Examples of Type Two Results include:

Child health conditions: Prevalence of children: who are born with low birth weights; who are fully immunized; who have functional limitations due to health conditions; who have age appropriate heights and weights; who are in good physical health, with no vision or hearing impairments.

Family income conditions: Prevalence of children: who live in poverty; who live with two parents or one parent employed.

Family life conditions: Prevalence of children: who are born to teen mothers or substance-abusing parents; who are abused; who live in foster care; whose TV viewing is regulated; who live in two-parent families; who live in low-crime neighborhoods.

TYPE THREE RESULTS—SERVICE-PROVISION AND ACCESS

Type Three Results are those that describe the services to which children and families have access. Distinct from the behaviors (Type One) or conditions (Type Two), this type focuses on *service provision and access to services that children and their families experience.* More than a tally of raw services, this type of results chronicles real availability of services to all ethnic, racial, and linguistic groups, with data typically reported in prevalences or percentages. Often Type Three Results include information about services by population sub-sets or individuals with particular conditions (e.g., disabilities, pregnancy, employed status). Data for these results are typically collected from record reviews and community and institutional data bases.

Examples of Type Three Results include:

Health provision/access: Prevalence of pregnant women who have access to early and continuing prenatal care. Prevalence of children: with increased access to prenatal care; with health insurance; who have access to regular vision and hearing screening, to medical care, to well-child examinations.

Parenting education provision/access: Prevalence of parents who have access to parenting classes and social supports.

Child Care/Preschool provision/access: Prevalence of low-income (or Limited English Proficiency [LEP] or disabled) children who have access to child care. Prevalence of children: who have access to developmentally appropriate child care and education; who have access to before and after-school care.

TYPE FOUR RESULTS—SYSTEMS CAPACITY

Rather than focusing on the provision of and access to discrete services, as indicated in Type Three Results, Type Four Results accord attention to *the way services are linked and function as a system.* Type Four Results assume that systemic capacity, efficiency, and integration are related to access and service quality which, in turn, are directly related to children's performance. Far less well developed than the other types, Type Four Results examine service redundancies, omissions, capacities, and efficiencies. Data for this type are collected in the aggregate and typically involve the amalgamation of information across agencies and service providers.

Examples of Type Four Results include:

Systemic efficiency: The degree to which the system uses its resources (e.g., fiscal, human, technical, and technological) efficiently and effectively.

Systemic infrastructure: The degree to which the infrastructure (e.g., training, financing, data gathering) supports efficient and effective service delivery.

Systemic accountability: The degree to which accountability is dispersed across systems; the degree to which agencies build collective accountability.

Systemic cultural sensitivity: The degree to which the system is sensitive and responsive to the needs of ethnically, racially, and linguistically diverse children and families.

Defining Different Purposes for Assessing Results

Different types of results exist because they are needed for different purposes. In some cases, for example, results information is needed by direct service providers for the purpose of enhancing the accuracy and quality of their work; in other instances, results data are needed for the purpose of demonstrating program efficacy; and in still other cases, results data are used to meet the information demands of policymakers and the public at large. Because these demands often exist simultaneously and because there has been no comprehensive data collection strategy delineated for young children, the purposes of amassing results data can and do become blurred. To address this confounding of purposes, various experts have proffered different schema (Bruner, Bell, Brindis, Chang, & Scarbrough, 1993; Shepard, 1995). All helpful, these schema have informed the following categorization of purposes for collecting results information. Each of the four purposes is addressed with respect to Type One Results information.

PURPOSE ONE—SCREENING AND EVALUATION

Information on Type One Results can be used for the purpose of locating children with specified characteristics, describing their current level of functioning, and determining their eligibility for intervention services. Typically, large numbers of children are quickly assessed to locate those few who might evidence a certain condition. The few, then, receive more thorough evaluation to learn whether some type of intervention is warranted. Formal and informal observations, checklists, tests, and parent interviews are commonly-used measurement approaches. Assessment for screening and evaluation may take place once or repeatedly.

Examples of assessments for this purpose include screening to discern the need for medical intervention (e.g., prescriptions, eyeglasses) or for special education services. In the latter case, the behavior of a single child is studied and compared with the average behavior of other children of the same age and characteristics (e.g., gender, geographic location). Large-scale assessments for this purpose include the Early Periodic Screening,

Diagnosis, and Treatment (EPSDT) program that funds the identification and treatment of health and developmental problems among Medicaid-eligible children, and Child Find a process to locate children who may be eligible for and benefit from special services, including those covered under the Individuals with Disabilities Education Act (IDEA).

PURPOSE TWO—IMPROVEMENT OF INSTRUCTION

Information on Type One Results can also be used for the purpose of providing feedback to teachers on the instructional process, with the intention of improving pedagogy, aiding in program planning, and creating learning experiences more appropriate to the needs of individual children. For this purpose, teachers may note the behavior of one child or a small group of children, using informal observations, checklists, anecdotal logs, or portfolios. Typically, such assessments occur on an on-going basis, and teachers need training to develop and hone their observation and assessment skills.

Examples of such assessments include teacher observation of children's level of small motor development in order to plan appropriate activities to foster such development or teacher observation of children's peer preferences and levels of play in order to arrange appropriate groupings of children.

PURPOSE THREE—PROGRAM EVALUATION

Assessment of children's performance and behavior for Purpose Three is undertaken to gauge the impact of a specific program or a particular intervention. The resulting data are likely to be used to guide future program design and funding decisions. For this purpose, the performance of groups of children is of interest, though children are likely to be assessed individually. Typically, such work is carried out by researchers, and data are reported about specific programs and interventions.

Examples of assessment for this purpose include the evaluation of the Parents as Teachers home visiting program or Kentucky's multi-age primary program.

PURPOSE FOUR—ACCOUNTABILITY

Children's knowledge and skills can also be measured for the purpose of informing the public about the collective status of children. For this purpose, the performance of children in classrooms, schools, districts, communities, states, and the nation is of interest; typically, progress is charted over time. For this purpose, groups of children are the unit for study, but not all children within a classroom or service unit will necessarily be assessed; samples of the group may be assessed. Assessment must be relatively time-efficient and the resulting data comparable and capable of aggregation.

Examples of results established for this purpose include parts of the Oregon Benchmarks, which are a series of results set by the public, that service providers, localities, and the state try to achieve and for which they are held accountable. One of these benchmarks is the percentage of children entering kindergarten meeting specific developmental standards for their age; communities are publicly challenged to improve the percentages of children achieving this result. In another example, Kentucky uses child results data to consider its allocations of state education funds. Assessment for accountability purposes usually has high stakes. Information about the findings tends to be broadly disseminated and used for decision-making. The highest stakes occur when recognition, funding, or other resources are directly tied to the reports of child performance.

Issues Concerning Definitions of Types and Purposes of Results

Although the definitions offered do render greater precision for the discussion, they also raise several key questions: What is the difference between inputs and results? Are all four types

really results? and What special issues arise when applying these constructs to very young children?

Across service spheres (e.g., education, health, social welfare) debate lingers regarding what constitutes an input or a result, and what distinguishes results from accomplishments. More than a semantic debate; the notion of what constitutes results warrants examination. Under many conditions, particularly in the education domain when speaking about students, "results" refer to what children know and can do, with inputs being the supports, materials, curriculum, pedagogy, and instruction that combine to foster the results. Alternatively, however, Type Two Results (child and family conditions)—while a means to student results—also constitute results in their own right. In short, there may be a chain of results (Types Two, Three, or Four) that lead to the ultimate goal of enhanced student performance (Type One). Some designate the results that lead to student results as interim results (Schorr, 1993); some consider them social indicators.

However thorny for children of any age, the questions of what constitutes inputs and results, and how to assess them, are particularly challenging regarding young children. Assessing young children's results is complex because of their episodic development, their dependence on adults and society for supports, and the disjuncture between the manner in which young children demonstrate competence (action and interaction) and more conventional approaches to measurement. As such, ethical questions emerge regarding the legitimacy of basing child results only on what young children know and are able to demonstrate. It is often argued that results for young children must be predicated on multiple types of data; Type One data alone are deemed too narrow an indication of child results. Information from all types, but most particularly Types Two and Three, must be included in an assessment strategy that takes full account of the age and expected abilities of youngsters. That is to say, when considering young children, conceptions of results evidence may need to be broadened to include what, for other age groups, may be considered inputs.

Chapter 3. *Desirability*

Service providers in early care and education hope to affect results for children—to prevent negative results from occurring and to promote positive results. Early childhood educators are quite used to on-going and informal assessment of young children. So while it might seem that there would be great receptivity toward child-based results in early care and education, this has not always been the case, because of the historical antipathy discussed in the introduction, the lack of training of many early care and education workers, and the lack of infrastructure in the early care and education system to collect results data. In addition, today's discussion of child-based results imposes new levels of rigor, specification, and accountability. Movement to a more formal and systematic child-based results approach would focus sustained public attention on the degree of attainment of specified results. Questions likely to be asked include: How effective are teachers? curricula? early care and education programs? schools? school districts? the amalgamation of services in communities, states, the nation?

Early care and education experts have divergent viewpoints on child-based results. One perspective is that it is unjust to predicate support for early care and education on the basis of results. Like education K-12, early care and education should be considered a *moral imperative* in a democratic society. This perspective also argues that the potential dangers of a child-based results approach outweigh the possible benefits. This perspective is articulated most emphatically under the following conditions: (a) the younger the children in question; (b) when using results data moves beyond the classroom; (c) when using results data to assess racially, ethnically, and linguistically diverse young children; and (d) when using results data to make high-stakes decisions about pay, program reimbursement, or other resource allocations. In this view, it may be desirable to assess children's results—particularly for children ages three through eight—for the purposes of screening and evaluating children (Purpose One), improving classroom instruction (Purpose Two), and perhaps for curriculum or program evaluation (Purpose Three). From this perspective, however, there are too many risks and not enough benefits to assess results for younger children and to aggregate data to use for accountability purposes in monitoring or decision-making in the community, state, or nation (Purpose Four).

Another perspective on results for young children is that the challenges of results definition and assessment—even for accountability (Purpose Four)—are addressable. This perspective regards the benefits of a results orientation as so attractive as to advocate immediate investments in constructing child-based results and systems for data collection for even very young children. The need for community, state, and national data to guide policy and practice, and even to allocate resources, is emphasized. This perspective argues that even if the early childhood education field delays, states and localities are preparing to assess child results, perhaps without needed advice from persons trained in child development and early education.

Amplifying these arguments, this section categorizes issues and then enumerates possible disadvantages and potential advantages of shifting to a child-based results approach for each. In most cases, specified disadvantages and advantages correspond to one or more of the indicated purposes. This section conveys the tensions surrounding child-based results, while later sections relate ideas for resolving them. In a paper commissioned for this Forum, Schorr reviews potential

benefits of an results-based approach, possible pitfalls, and promising strategies for implementation (see Appendix D).

The Impact on Teachers' Practice and Children's Experiences

What effect, if any, would child-based results have on daily practices affecting children and their teachers? Those who oppose a results-based approach feel that it would direct teachers to teach to the test, thereby limiting their creativity and the spontaneity and flexibility of early education. Advocates for a results-based orientation believe that it would help to guide practice, making it more purposeful and goal driven. These positions are elaborated more fully below.

Potential *advantages* for teachers' practices and children's experiences:

- Teachers would have more information about children's learning and individual differences, allowing teacher practices to address children's individual needs and backgrounds, and expanding practices to address multiple areas of child development rather than just cognitive skills and knowledge (Purposes One, Two, and Four)
- Teachers would develop and increase effective instructional approaches, curricula, and services if they know more about what works, have the flexibility to design approaches rather than being required to use prescribed methods, and have precise goals toward which to teach (Purposes Two, Three, and Four)
- Teachers would use information on child development to communicate more effectively with family members, to help them support their children's learning (Purposes Two and Four)
- Teachers would have increased expectations for all children—particularly ethnically, racially, and linguistically diverse young children—if all children are expected to achieve the same high results (Purpose Four)

Possible *disadvantages* for teachers' practices and children's experiences:

- Teacher practices would become less effective if the results are misleading and do not capture the nature, complexity, and individuality of children's development—particularly for ethnically, racially, and linguistically diverse young children (Purposes Two and Four)
- Teacher practices would become more uniform, in the attempt to achieve uniform results; homogeneous services could not meet the unique needs of individual children (Purposes Two and Four)
- Communication with family members would become less useful—emphasizing performance on test items rather than overall child development (Purposes One, Two, and Four)
- Children likely to test poorly and lower school averages would be retained or their entry to school delayed, particularly minority children and children with special needs or limited English proficiency; as a result, children could be labeled or stigmatized (Purposes One, Two, and Four)

The Impact on Public Understanding

The desirability of a results approach also depends on how the data are interpreted by families and the public. If the results are narrow, trivial, abstract, or culturally-bound, then a results approach might hinder public understanding. If the results are meaningful to families and the public, and if they are expressed in everyday language, then a results approach might serve to instruct the broader society about child development.

Potential *advantages* for public understanding of young children's development:

- Public knowledge and general understanding of healthy child development and of developmental problems would increase if results are meaningful and valuable to parents and the public (Purpose Four)
- The public would come to understand more about the multi-dimensional nature of child development; understanding of the relationships among children, families, services, and systems of the early years would increase (Purpose Four)
- Appropriate assessments would lead to the development of realistic expectations for the development of children and the performance of the programs in which they participate (Purpose Four)

Possible *disadvantages* for public understanding of young children's development:

- The public would draw invalid conclusions about children's abilities from results that are inappropriate for all or some young children. For example, results that focus on cognitive skills minimize the importance of other dimensions of early learning. Another example would be an assessment system that does not reflect how the performance of low-income children may have improved over time (Purpose Four)
- The public would think that child development is simpler and more uniform than it is, because it is impossible to capture the complexity and individuality of development in specific results (Purpose Four)
- The public would be confused about what helps and hinders learning, if data about families, communities, services, and systems are not collected and presented as the context for child results (Purpose Four)
- The public would place far more pressure on young children by developing false, unrealistic expectations for performance (Purpose Four)

The Impact on Funding

Current discussion about the desirability of adopting a results orientation in early care and education occurs within the framework of the devolution of government responsibility, program consolidation, budget cuts, and heightened competition for funds. Those who question a results-based approach are concerned that it will lead to a reduction in funding for services for children and families in general, for early care and education, and for low-income and minority children who may not perform well on tests. Advocates for a results-based approach feel that having dependable data is the only hope for maintaining—and possibly increasing—funding levels and services.

Potential *advantages* for funding:

- Policymakers or government administrators would use results-based data to reallocate funds from marginal and ineffective early care and education programs to effective programs (Purpose Four)
- Program administrators would use results-based data to expand more effective programs and improve or eliminate less effective ones (Purpose Four)
- Investment would increase in services to low-income and otherwise needy children, in prevention efforts (rather than remediation), and in infrastructure, particularly if the cost savings of this additional funding is documented (Purpose Four)

Possible *disadvantages* for funding:

- Effective early care and education programs would have their funding cut if the chosen results are insignificant, narrow, or insensitive to family and community differences. (Purpose Four)
- All early care and education programs would have funding cut if results are not met (whether or not these programs are at fault)

and if the public loses hope or becomes cynical (Purpose Four)

- There would be fewer resources to provide early care and education services to children if funds are diverted to setting, assessing, interpreting, and communicating results (Purpose Four)
- What limited infrastructural support that currently exists would be threatened because of the desire to keep resources close to the children so that results will improve (Purpose Four)

The Impact on the Relationship of Early Care and Education and Other Services

As society fails to solve the complex human problems in today's communities, there is a trend toward services integration and an increasing acknowledgement that the comprehensive needs of families cannot be met with narrow, categorical services. Those who question taking a results orientation feel that it might foster competition among all social services and economic development; fragmentation among services will grow as competition increases. Proponents of a results-orientation believe that it will provide the vehicle for varied social service and economic development efforts to work together toward improving the lives of children and families.

Potential *advantages* for the relationship of early care and education with other services:

- Agencies and programs across service areas would cooperate, collaborate, and possibly combine funds to achieve positive results; planning across social services would be facilitated by results data (Purpose Four)
- The gap between early care and education and elementary education would be bridged with the common focus on results, as might the rift between services for children with special needs and education in general (Purpose Four)

Possible *disadvantages* for the relationship of early care and education to other services:

- Early care and education programs would lose resources and attention if other services have better results (Purpose Four)
- Ill will and fragmentation among services would increase and collaboration decrease if some services have better results (Purpose Four)
- The various service sectors would blame each other for poor results if attribution of results is not demonstrated clearly (Purpose Four)

So ... What Do We Conclude About The Desirability of Moving to a Child-Based Results Approach?

Returning to the possible purposes of a child-based results orientation, there is some consensus in the early childhood field that gathering information for preschool- and primary-aged children for screening and evaluation (Purpose One) and for improvement of instruction (Purpose Two) is advantageous for children and families, provided that the results chosen are both significant and appropriately measured. Many early childhood experts also favor the use of child results for program evaluation (Purpose Three). There is less agreement regarding movement to child-based results for these purposes for children below the preschool years, notably infants and toddlers.

While some first Forum participants strongly supported moving to child-based results for accountability purposes, there was little agreement about the wisdom of doing so for all young children, birth to age eight. Even among those who support such movement, there is greater

support for assessing child-based results for monitoring and planning than for using results to allocate resources.

To implement a child-based results approach that avoids the possible disadvantages described above and achieves the advantages enumerated, planners must meet certain necessary conditions. Implementation of a results orientation is desirable *only if* critical safeguards are in place during the process of results identification, while planning the assessment, throughout the data collection, and as findings are interpreted and communicated. These "only ifs"—or necessary conditions—are discussed in the section that follows.

Chapter 4. *Feasibility*

Despite disagreement on the desirability of moving to a child-based results approach for accountability purposes, many experts believe that doing so now is more feasible than at any other time in our history. They indicate that certain fields, such as early childhood special education, have focused on child results for more than 20 years and have both good-practice and bad-practice examples to inform the discussion. Further, current research is improving the options for gathering information that is both valid at the time of measurement and meaningful across the developmental course. In papers commissioned for this Forum, White explores some of the limits of existing standardized tests for young children (see Appendix C) and Love presents arguments for the feasibility of a child results approach (see Appendix E). Despite these advances, all are concerned that any shift to a results orientation be made in ways that are developmentally, culturally, and contextually appropriate. Care must also be exercised in the process of developing and interpreting results. Inherently precarious and high stakes, the effort to develop and implement a results approach can only be undertaken under certain conditions. Moving to child-based results is feasible "only if" the following five conditions are met:

"Only If" There Is Broad Participation In The Identification of Results

INCLUDE MANY STAKEHOLDERS IN RESULTS DEVELOPMENT

Results that will be useful to the nation need to be agreed upon by a broad constituency, including parents, policymakers, practitioners, and researchers. Politicians, government administrators, business leaders, and citizens have meaningful contributions to make in the development of results as do individuals from diverse ethnic, racial, and linguistic backgrounds. The process of building consensus by selecting certain child results that are significant to broad audiences is crucial. Moreover, in developing results, it is important to remember that the input of lay citizens can help assure that results are meaningful and locally appropriate.

WORK WITH PRACTITIONERS AND PARENTS IN PARTICULAR

Given the discomfort of many in the early childhood education community (personnel along with parents) regarding a shift to a results orientation, it is important that conversations regarding child-based results involve practitioners and parents. Personnel in early care and education need to converse with others in the education and human service fields who are already using a results approach—such as early childhood special educators, elementary school teachers, and health care providers—to gather suggestions from their experiences. Parents need to understand clearly the implications of moving to a child-based results orientation. More significantly, both groups know children well; their ideas are critical to constructing effective, appropriate results.

"Only If" We Can Identify Appropriate Results

CHOOSE RESULTS THAT ARE SIGNIFICANT FOR CHILD DEVELOPMENT AND THE LIFE COURSE

As a child-based results approach evolves, there may be a tendency to choose results that can be easily measured or to select "quick and dirty" indicators for which measures already exist. In lieu of these sometimes flawed approaches, plan-

ners need to consider what results they deem fundamentally important, with major findings from child development influencing the content of the results. Good results have merit because they are important in and of themselves, and because they are linked to longer term life goals. The discussion of life-course results is also valuable because it can unite Americans across racial and ethnic lines; most people want the same major life-course results for their children. Thus, the essential task of results planners is to take the complex constructs that research has demonstrated to be important (e.g., identity formation, achievement motivation, task persistence, establishing peer relationships) and identify the life-course "edge" appropriate to the age group under consideration.

Life course results might include being ready for school, being able to read, graduating from high school, attending college, holding a job, or avoiding teenage pregnancy and crime. Conceptualized in this way, such results will have utility for parents and policymakers. Moreover, they contextualize the content of the results in a more durable, long-term perspective that is salient across all populations. For these reasons, life-course results should be considered.

CHOOSE RESULTS THAT CROSS MULTIPLE DOMAINS OF CHILD DEVELOPMENT

Basing educational results on subject matter curriculum areas (e.g., science, math), while perhaps appropriate for older children, needs to be examined critically when considering younger children. Learning for young children is less oriented to subject matter facts than to the fostering of basic developmental competence. To that end, the Goal 1 Technical Planning Group of the National Education Goals Panel, building on decades of work by scientists and practitioners, has identified five dimensions of early learning that provide a developmental, rather than a curricular, framework. The dimensions are: physical well-being and motor development; social and emotional development; approaches toward learning; language usage; and cognition and general knowledge (Kagan, Moore, & Bredekamp, 1995). Results for young children need to consider these dimensions, as well as a curricular orientation.

Beyond incorporating a broad-based developmental orientation, results for young children must take into account children's unique learning approaches. Young children, especially, do not learn in compartmentalized categories; they amass knowledge through integrated experiences. Consequently, results for young children must reflect integration across domains and across subject areas. Results for young children should not focus only, for example, on cognitive development, but must emphasize all domains. It is imperative that the domains be considered as a totality, with no "single domain acting as a proxy for the complex interconnectedness of early development and learning" (Kagan, Moore, & Bredekamp, 1995).

Use of multi-dimensional results will also minimize teaching aimed solely at producing high test scores. For example, it is easy to teach a child to label pictures of community helpers (single-dimension skill), but more challenging to teach children how to plan and play with others (multi-dimensional skill).

CHOOSE RESULTS TO WHICH DOLLAR VALUE CAN BE ATTACHED

Given the precarious funding of early care and education and given the persuasiveness of cost-saving data to policymakers, new results should be amenable to cost analysis. To date, in a very limited number of studies, the early care and education field has been particularly successful in demonstrating the cost-effectiveness of early intervention. Recognizing these conditions, new results data must be responsive to policymakers' thirst for additional fiscal information.

CHOOSE RESULTS THAT REFLECT THE REALITIES OF EXISTING COMMUNITIES AND PROGRAMS, YET CAN BE AGGREGATED

Real world circumstances must guide the adoption of results. If schools in a community teach only in English, then an appropriate result must be that children read, write, and speak English. On the other hand, if children in a community are allowed to demonstrate their competence in Spanish, Korean, or English, then the results process must reflect that fact. In communities in which neighborhood culture and school expectations differ, results might reflect children's ability to function successfully in both settings. Selected indicators must not gloss over the differences in "readiness for kindergarten" in the myriad of communities across America, nor can local values be ignored. Nevertheless, if results are to be used to monitor local, state, and national trends, a core group of results must be drawn to allow comparisons among programs, communities, and states.

DETERMINE WHETHER RESULTS SHOULD BE COMPLEX OR SIMPLE

One perspective is that results must reflect the complexity of development; indicators stripped of developmental "richness" are not meaningful. Another view is that worthy child results can focus on a few crucial elements; these would be stage-salient tasks with considerable social validity, such as trust behaviors in infancy or reading at the end of first grade.

The issues of the complexity and "developmental embeddedness" of selected results, along with the intricacies of the assessment methods adopted to measure them, determine the costs of these approaches. One perspective underscores pragmatism in a time of tight budgets; namely, early care and education should move forward to develop quick, low-cost approaches to results assessment. Another view is that simple measures of complex developmental phenomena are not presently possible. Given that fact, the field should honestly explain its position and inform policymakers that adequate time and money must be provided before results assessment can go forward.

In either case—complex or simple—the results should be designed to "tell a story" to increase the understanding of the public and policymakers of child development.

"Only If" We Are Clear About Which Children To Include

DETERMINE HOW TO INCLUDE CHILDREN WITH LIMITED ENGLISH PROFICIENCY AND SPECIAL NEEDS

To what extent should children with limited English proficiency be included in results assessments? A predominant view is that all children should be included in such a system. Otherwise, the system would not be national; it would not be just. However, with children whose home language is other than English, and particularly in instances where multiple languages are represented in a classroom or community, assessment technicalities and practical issues exist.

A similar issue arises with regard to young children with special needs. In several states, youngsters with special needs have been "overlooked" in designing results assessment systems, while other states have included them. Recommendations from national special education experts (National Center on Educational Results, 1993), and pioneering results efforts in several states, suggest that most children with disabilities should participate in results evaluations.

"Only If" We Measure Results Appropriately

LOOK AT CHANGES FROM BASELINE BEHAVIOR OVER TIME

Because young children's growth is highly episodic and variable, performance cannot be judged at a single point in time, but must be gauged from repeated observations; data must be collected at multiple points in time. Use of measures over time also reveals developmental progress in children whose exposure to early learning opportunities has been restricted and whose baseline and current performance are delayed for their chronological age. In monitoring local programs, states, and the nation, it is essential to realize that improvement in results must be regarded relative to children's starting points. This is particularly important when studying children at risk, who may be making great gains but still have below average results. Results assessments take into account starting points and must look at change over time.

FOCUS ON FACE VALIDITY, CONSTRUCT VALIDITY, AND CONSEQUENTIAL VALIDITY

The paragraphs above have underscored the importance of results and results measures that "make sense" to parents, the public, and practitioners. This is the concept of face validity. Construct validity is also critical: Do the assessment approaches measure what they are intended to measure? Do they clarify performance consistent with the concept embodied in the results? Consequential validity is also very important: Are results used in valid ways (i.e., to make accurate explanations)? Is the interpretation given consistent with the actual findings?

THINK INVENTIVELY ABOUT ASSESSMENT

A great deal of innovative assessment is taking place nationally, and new results efforts should seek to incorporate this cutting edge work. Among these efforts, play-based assessment, classroom observation schemes, authentic assessment, descriptive documentation, interviews with teachers and parents, portfolio approaches, and Vygotskian dialogues show promise for providing useful developmental information. While each of these approaches faces unique problems, they share the challenge of aggregating descriptive information in a concise form that can be used in decision making. As such, they provoke thinking about critical issues that need to be faced as results information is generated and made useable.

"Only If" We Link Child-Based Results to Efforts to Improve The Lives of Children

LINK CHILD-BASED RESULTS WITH OTHER INDICATORS

If a purpose of results measurement is to help practitioners, parents, policymakers, and the general public understand the status of young children, and to improve this status, then it is essential that the conditions in which children are living and learning (Type Two) be assessed and related to the child-based results. Moreover, the services that children are receiving—or not receiving—(Type Three) and some or all of the elements of the systems that support early care and education must be assessed and related to the child-based results.

Information about multiple types of results should be used to help the public understand the circumstances upon which child results depend. Some early childhood experts are more vocal than others in requiring this linkage in any results system that is developed. More data are already

being aggregated for Types Two and Three than for child-based results (Type One), and, at the present time, there is little reporting of child results along with the other types of information.

Chapter 5. *Suggested Next Steps*

While significant progress has been made in delineating the issues surrounding the use of results in early care and education, the topic remains conceptually complex, practically challenging, and politically sensitive. The suggested next steps that follow build upon preliminary discussion from the first Forum and elaborated discussion at the second Forum. The next steps embody a number of recommendations for advancing a results-based approach, but they are intentionally suggestive, offering domains and strategies for action that need to be honed.

Increase Public Consciousness and Participation

CONSIDER TERMINOLOGY

Within discussions of results, similar terms often convey different meanings to various speakers; alternatively, different terms may convey the same concept. At the same time, particular terms may be politically charged in certain locations with specific audiences, but not in others. In short, there is no clear set of terms with which to conduct the debate. Since language is the vehicle for meaning, foundational concepts need to be clearly stated and mutually agreed upon early in the discussion of child-based results in early care and education.

BROADEN PARTICIPATION IN IDENTIFYING RESULTS AND BUILD CONSENSUS REGARDING THEM AT LOCAL, STATE, AND NATIONAL LEVELS

Worthwhile results that are broadly "owned" can result only from shared construction by parents, teachers, administrators, researchers, policymakers and the public at large. Building consensus at the local, state, and national levels on desired results that are meaningful to varied audiences is critical to raising public consciousness and supporting an ongoing effort. An intentional focus on including traditionally disenfranchised groups in the process is essential. All consensus-building efforts need to be coordinated by groups considered to be non-biased and legitimate.

ENGAGE THE EARLY CHILDHOOD COMMUNITY

The early childhood community is correctly concerned about the development and implementation of results for very young children. The concerns of this community need to be fully understood and carefully addressed. Forums for early childhood practitioners and researchers need to be held; written materials need to be developed. Above all, time needs to be allowed for support and consensus to emerge.

Plan Strategically

LEARN FROM AND BUILD ON EFFORTS IN CHILD DEVELOPMENT AND ALLIED FIELDS

Given the focus on results creation, construction, and collection throughout the nation, it seems wise to assess and utilize—where appropriate—the data currently being collected for other efforts. In particular, a number of government and foundation projects have developed models for child-based educational results for younger and older children. Data gathering efforts for older children, including the National Assessment of Educational Progress, and for younger children, such as the Early Childhood Longitudinal Study, have struggled with of the same issues explored in this paper that challenge efforts to move to a child-based results orientation. It would be especially informative to consider how they have—and have not—addressed the necessary conditions (only ifs).

Other sources of guidance are child-based

results efforts in related fields, such as child welfare, adoption, foster care, early childhood special education, and child health. Attention to the successes and failures in results approaches across disciplines will be invaluable at the formative stage of efforts in early care and education. Additionally, others in early care and education have pondered these same issues. They include the Goal 1 Technical Planning subgroup and Head Start Research Committees. Efforts should be made to use the expertise and documents from these groups.

CONSIDER AND PREPARE FOR UNINTENDED CONSEQUENCES

Moving to a results orientation will yield some unintended consequences. To the extent possible, efforts to predict such consequences, particularly those that might be negative, are encouraged. In addition, once potential negative consequences are identified, strategies to deal with them need to be developed and implemented.

COORDINATE RESULTS-RELATED EFFORTS

Create a collaborative of organizations engaged in work on results related to early care and education to support and inform one another. Such a group could not only cross-fertilize existing work and minimize duplications, but could also create strategic plans delineating where additional technical, consensual, and political work is needed.

Identify and Choose Results Carefully

IDENTIFY RESULTS BY TYPE

Confusion exists regarding different types of results. To that end, different types of results should be discerned. For results related to children's behavior, a broad national consensus needs to be developed, with ample opportunity for states and locales to tailor their specific results and benchmarks. Such results should be strengths-based and should allow for the reflection of partial achievement. Efforts to specify Type One results might look to special education which has been using similar results for many years. For results related to child and family conditions (Type Two), access and quality of services (Type Three), and systemic results (Type Four), model results could be developed at the national level for state adoption. All results should evolve and be subject to frequent change as knowledge, social conditions, and values are altered.

IDENTIFY LIFE-COURSE RESULTS

Define major life-course results for older children, and specify the antecedents in early childhood that symbolize important "real life" skills. Tying life-course results to younger children will help to elevate the importance of the early years.

IDENTIFY LINKS AMONG THE RESULTS TYPES

Because results are interrelated, clearer pathways among and between individual results and results types need to be clarified. In particular, linkages need to made between results related to children's behavior and knowledge (Type One) and the context in which children develop, as expressed in results related to child and family conditions (Type Two), service provisions (Type Three), and systemic integration (Type Four).

FOCUS ON POSITIVE RESULTS

Frequently, particularly regarding child and family conditions (Type Two results), there has been a tendency to chronicle negative results. Efforts should be made to identify and assess positive results (e.g., resilience, protective factors, child well-being), as well.

SELECT IMPORTANT RESULTS TO WHICH TEACHERS CAN AND SHOULD TEACH

The results chosen for young children must be truly important to their future development and achievement, not simply indicators that are easy to measure. When results are salient and when they are evaluated in contextually-sensitive ways, teachers can conduct activities that promote the skills that assessments measure.

RESOLVE THE TENSION BETWEEN THE DESIRE OF COMMUNITIES AND STATES TO CUSTOMIZE RESULTS AND THE NEED TO AGGREGATE DATA

It is necessary to identify results that are meaningful to local communities and states to inform the process and assure local ownership of the results. It is also crucial to maintain some consistency across all data such that they are useful, interpretable, and comparable.

AIM HIGH, BUT BE REALISTIC

Child-based results must be selected that create high expectations for all children. Such well-designed results will guide service improvement and policy formation. Care must be taken, however, not to over-reach or idealize the conditions under which children live or the services they can reasonably be expected to receive.

Develop Appropriate, Cost-effective Approaches to Assessment and Data Collection

DEVELOP MEASUREMENT TECHNIQUES FOR LARGE SAMPLES OF YOUNG CHILDREN

Efficient methods for gathering data from children in a variety of settings must be pioneered, building upon the experiences of multi-site early childhood education studies conducted within the past decade. As noted earlier, promising new techniques (e.g. play-based assessment, classroom checklists for anecdotal records, and Vygotskian dialogue methods) merit exploration because they reflect child development and the context in which it occurs. Narrow, multiple-choice psychometric measures are not acceptable for any of the purposes herein. Richer more complex data show promise of being useful for multiple purposes. Worthy goals are to collect fewer data and make the best possible use of what is collected, and to give children from diverse backgrounds the opportunity to perform to their optimum capacity.

DEVELOP ACCEPTABLE APPROACHES TO INCLUDE CHILDREN WITH LIMITED ENGLISH PROFICIENCY AND THOSE WITH SPECIAL NEEDS

Several states and organizations have developed approaches to including as many individuals as possible in results assessment. To do this, results measures and assessments need to be translated into multiple languages, appropriate for multiple cultures, and accessible by children with special needs.

Put Theory Into Practice

PROVIDE TECHNICAL ASSISTANCE TO SUPPORT RESULTS DEVELOPMENT AND ASSESSMENT

States are moving to using child-based results rapidly. There is an urgent need to support state and local practitioners as they surge forward. In many cases, states will be pressed to implement a results system well before many of the issues can be addressed fully. To help states in the meantime, information across states should be chronicled and mechanisms established for information sharing among those developing results and related assessment systems.

PROVIDE TECHNICAL ASSISTANCE TO SUPPORT DATA GATHERING AND MEASUREMENT

Presently, many states are enhancing their data gathering and measurement capacities. Not only are states using approaches that differ from one another, but often administrative departments within states differ in their approaches to data gathering and measurement, preventing the aggregation of data germane to specified results. To that end, technical assistance should be provided to foster comparable measurement and assessment approaches across state administrative agencies and perhaps across states.

PILOT WELL-CONSTRUCTED MODELS

Any results system for early care and education must be grounded in accepted child development theory and validated assessment practices. For collecting results-based data for accountability purposes, it is crucial that innovative approaches from a variety of disciplines and audiences be considered. Before policies are drafted, approaches should be pilot-tested at several diverse sites to assure the efficacy of the approach. Expansion should proceed at a manageable pace.

CONTINUE THE DIALOGUE

Convene meetings once or twice a year to review results work being done by districts, states, and individual researchers; to identify exemplary efforts and find ways to share them widely; and to identify gaps in knowledge and plan ways to fill them. Future gatherings might continue to explore divergent viewpoints on difficult issues and spark inventive resolutions.

Explore Ways to Fund a Results Approach Adequately

DETERMINE NECESSARY FUNDING

The collection of some data may require no additional funds. For example, funding for data collection around special education placement is available, and funding for research on instructional approaches may be contained within the development and pilot testing budgets for such projects. However, monitoring child results for accountability purposes is typically a project that needs specific funds. A necessary first step is to define the scope of the proposed project, noting the complexities and costs inherent in large-scale data collection.

EXPLORE POTENTIAL FUNDING SOURCES

Both start-up and ongoing funding need to be explored, with consideration given to "piggybacking" with other data collection wherever possible. Funding options to be considered include foundations, government agencies, and state agencies.

Communicate, Implement, and Evaluate

CONCENTRATE ON COMMUNICATING RESULTS BROADLY AND EFFECTIVELY

To have impact, results data on young children must be shared with parents, policymakers, business, and the media. Careful consideration must be given to the nature of the data shared and the process for sharing it. Not all data are in a form immediately suitable for use by multiple audiences. Care must be taken to assure effective communication to appropriate audiences.

CONTINUOUSLY EVALUATE AND IMPROVE RESULTS MEASURES AND APPROACHES

Initially, no process of results and assessment will be complete or ideal. Efforts and instruments need to be monitored and evaluated continually. Realistic timelines and comprehensive and sequenced plans should be developed so that results-based efforts are regularly re-examined.

REFERENCES

Bruner, C., Bell, K., Brindis, C., Chang, H., & Scarbrough, W. (1993). *Charting a course: Assessing a community's strengths and needs.* Falls Church, VA: National Center for Service Integration.

Graue, M. E. (1993). *Ready for what? Constructing meanings of readiness for kindergarten.* Albany, NY: State University of New York Press.

Gnezda, M. T., & Bolig, R. (1989). *A national survey of public school testing of prekindergarten and kindergarten children.* Paper commissioned by the National Forum on the Future of Children and their Families of the National Academy of Science and the National Association of State Boards of Education.

Institute for Research on Poverty. (1995, Spring). *Focus,* 16(3).

Kagan, S. L. (1995). *By the bucket: Achieving results for young children.* Washington, DC: The National Governors' Association.

Kagan, S. L., Moore, E., & Bredekamp, S. (1995). *Considering children's early development and learning: Toward shared beliefs and vocabulary.* Washington, DC: National Education Goals Panel.

Love, J. M., Aber, L., & Brooks-Gunn, J. (1994). *Strategies for assessing community progress toward achieving the first National Educational Goal.* Princeton, NJ: Mathematica Policy Research, Inc.

Meisels, S. (1988). Developmental screening in early childhood: The interaction of research and social policy. *Annual Review of Public Health,* 9, 527-550.

National Center on Educational Results. (1993). *Educational results and indicators for early childhood (Age 3 and Age 6).* Minneapolis, MN: University of Minnesota, College of Education.

Phillips, D. & Love, J. (1994). *Indicators for school readiness, schooling, and child care in early to middle childhood.* Paper prepared for the Indicators of Children's Well-Being Conference, November 17-18, Rockville, MD.

Schorr, L. B. (1993). *A framework for improving results for children and families.* Washington, DC: Improved Results for Children Project.

Schorr, L. B. (1994). The case for shifting to results-based accountability. *In Making a difference: Moving to outcome-based accountability for comprehensive service reforms* (pp 14-28). Falls Church, VA: National Center for Service Integration.

Shepard, L. (1995). The challenges of assessing young children appropriately. *Phi Delta Kappan,* 76(3), 206-212.

Shepard, L. A., & Smith, M. L. (1987). Effects of kindergarten retention at the end of first grade. *Psychology in the Schools,* October, 346-357.

Young, N., Gardner, S., & Coley, S. (1993). Getting to results in integrated service delivery models. In N. Young, S. Gardner, S. Coley, L. Schorr, & C. Bruner (Eds.), *Making a difference: Moving to outcome-based accountability for comprehensive service reforms.* Falls Church, VA: National Center for Service Integration.

Zero to Three. (1992). *Heart start: The emotional foundations of school readiness.* Washington, DC: Author.

Appendix A: *June 1–2, 1995*

Meeting Agenda and Participants

FIRST ISSUES FORUM ON CHILD-BASED RESULTS AGENDA

GOAL: To examine the *desirability* and *scientific feasibility* of moving toward a child-based results orientation for children birth to age eight.

June 1 and 2, 1995
Carnegie Corporation of New York
437 Madison Avenue, New York, NY

Day One
June 1, 1995

11:00 Buffet brunch available

Session I — Welcome and Overview
Sharon L. Kagan, Chair

12:00 Welcome/Introductions
Goals of the meeting
Background, Rationale, and Definitions

12:45 Considerations Regarding an Results-Based Orientation
Presentation by Lisbeth B. Schorr

1:05 Considerations Regarding an Results-based Orientation
Presentation by Sheldon White

1:25 General Group Discussion

2:45 Break

* * * * *

Session II — Implications for Children's Development
Michael Levine, Chair

3:00 Participants will be asked to respond to the following questions:

a. What are the necessary characteristics of an assessment process that obtains needed information and also promotes child development? How does the process differ depending upon whether we are assessing for instructional improvement or accountability?
b. Are there special conditions or needs of young children in general, and certain young children in particular, that might exempt them from or prefer them for inclusion in results assessment?
c. How can a results orientation be structured to be fully comprehensive for children, 0-8?

4:00 General Group Discussion

5:00 Adjournment

* * * * *

Session III — Dinner and Roundtable Discussions on Desirability
Valora Washington, Chair

6:30 Cocktails

7:00 Dinner

7:45 Concurrent Roundtable Discussions

Roundtable A — Impact on Classroom Practice and Teacher Preparation

a. What is the potential impact of a results orientation on classroom practice? on professional development?
b. If a results approach were adopted, what special actions should be taken to encourage positive results and prevent negative consequences in classroom practice?

Roundtable B — Impact on Programs

a. How would a movement toward child results impact program quality? availability?
b. What is the relationship between results and funding?
c. What program policies might be altered (positively or negatively) as a result of a results-orientation (e.g., retention)?

Roundtable C — Impact on Families and Communities

a. How will parents and diverse community groups respond to a results orientation (e.g., persons with low income, business leaders, service providers)?
b. What precautions are necessary to prevent misuse of outcome data within a community?

Roundtable D — Impact on Early Childhood Systems

a. Would the benefits and challenges of a results approach touch all segments of the field evenly? Which constituencies/programs would be most affected? How?
b. Would movement toward a results orientation help unite or further divide early care and education?
c. How would a results approach affect monitoring and licensing? advocacy?

Roundtable E — Impact on State and Federal Policy

a. What are the potential policy consequences of a results orientation in the states? at the national level?
b. If a results approach were to be adopted, what strategies should be implemented at the national and state levels to ensure maximum benefits and constrain harm in the states?

9:00 Adjournment

* * * * *

Day Two
June 2, 1995

Session III Continued — Discussion on Desirability
Valora Washington, Chair

8:00 Continental breakfast

8:30 Reports from Roundtable Discussions
A — Impact on Classroom Practice and Teacher Preparation
B — Impact on Programs
C — Impact on Families and Communities
D — Impact on Early Childhood Systems
E — Impact on State Policy

9:20 General Group Discussion

10:20 Break

Session IV — Scientific Feasibility of an Results Approach
Michael Levine, Chair

10:35 Review of Past Efforts and Scientific Feasibility
Presentation by John Love

10:55 Participants will be asked to address the following questions:
a. What cultural, technical, developmental, and implementation considerations must be kept in mind if a child-based results orientation were to take root?
b. Do we understand the pitfalls and can we overcome them, now?

12:00 General Group Discussion

12:45 Lunch

1:30 Participants will be asked to respond to the following questions:
a. For each age group (infancy, preschool, and primary), is the development and implementation of child-based results technically feasible at this time?
b. What, if any, are the special considerations for your age group?

2:15 General Group Discussion

3:15 Break

* * * * *

Session V — Summary and Conclusions
Sharon L. Kagan, Chair

3:30 Summation: The Desirability and Scientific Feasibility of an Results Approach
Presentation by Barbara Blum

4:00 General Group Discussion: Next Steps

4:30 Adjournment
First Issues Forum on Child-Based Results

June 1–2, 1996
Participants

Larry Aber
National Center for Children in Poverty
Columbia University

Barbara Blum, President
Foundation for Child Development

Sue Bredekamp
National Association for the Education of Young Children

Cynthia Brown
Council of Chief State School Officers

Charles Bruner
Child and Family Policy Center

Bettye Caldwell
Pediatrics/CARE
Arkansas Children's Hospital

Nancy Cohen
Bush Center in Child Development and Social Policy
Yale University

Ellen Galinsky
Family and Work Institute

Sarah Greene
National Head Start Association

Kenji Hakuta
School of Education
Stanford University

Sharon L. Kagan
Bush Center in Child Development and Social Policy
Yale University

Mary Kimmins
Maryland State Department of Education

Luis Laosa
Educational Testing Service

Michael Levine
The Carnegie Corporation of New York

Joan Lombardi
Administration for Children, Youth, and Families
U.S. Department of Health and Human Services

John Love
Mathematica Policy Research, Inc.

Samuel J. Meisels
University of Michigan

Kristin Moore
Child Trends

Frederic Mosher
The Carnegie Corporation of New York

Deborah Phillips
National Research Council

Craig Ramey
Civitan International Research Center
University of Alabama at Birmingham

Sharon Rosenkoetter
Bush Center in Child Development and Social Policy
Yale University

Lisbeth Schorr
Harvard University Working Group on Early Life

Diana Slaughter-Defoe
School of Education and Social Policy
Northwestern University

Robert Slavin
Johns Hopkins University

Valora Washington
The W.K. Kellogg Foundation

David Weikart
High/Scope Educational Research Foundation

Charles E. Wheeler
Walter R. McDonald & Associates, Inc.

Sheldon White
Department of Psychology and Social Relations
Harvard University

Emily Wurtz
Office of Senator Jeff Bingaman

Nicholas Zill
Westat, Inc.

FOUNDATION REPRESENTATIVES

Stacie Goffin
Ewing Marion Kauffman Foundation

Mary Larner
David and Lucile Packard Foundation

Janice Molnar
Ford Foundation

Appendix B: *January 24, 1996*

Meeting Agenda and Participants

Second Issues Forum:
Next Steps for Child-Based Results
Agenda

Wednesday, January 24, 1996
10:00 am to 4:30 pm
Carnegie Corporation of New York
437 Madison Avenue, New York, NY

Goal: To identify "actionable" next steps for developing a results-oriented approach for accountability in early care and education

10:00 to 10:15 Welcome, Introductions, and Charge to the Group

10:15 to 10:30 Questions and Discussion Concerning First Meeting

10:30 to 12:00 Identifying Results

Is there a need to identify and build consensus on key type 1 results (what children know and can do) or have such results been adequately identified? Consider this question separately for children age 5, age 3, and younger than age 3. What are the next steps in this area?

Is there a need to identify and build consensus on key type 2 results (child and family conditions), or have such results been adequately identified? Consider this question separately for children age 5, age 3, and younger than age 3. What are the next steps in this area?

Is there a need to identify and build consensus on key type 3 results (service provision, access, and quality) for families and children, or have such results been adequately identified? Consider this question separately for children age 5, age 3, and younger than age 3. What are the next steps in this area?

What is meant by type 4 results (systems capacity)? How do they relate to type 1, 2, and 3 results? Is it necessary to develop different type 4 results for children age 5, age 3, and younger than age 3? What are the next steps in this area?

(If time allows) What are the "markers of progress" for results types 1-4? What are the micro-results that lead to other micro-results? What are the complex relationships among the different types of results? Which results lead to which other results? What are the next steps in this area?

12:00 to 1:00 Assessing Results

Is additional work on instrument development and assessment methods needed for implementing a child-based results approach? If yes, for which types of results and for children of which ages? Are there adequate approaches for assessing some relatively complex results (e.g. nurturing families)? What are the next steps in this area?

Have cost-effective approaches to data gathering and assessment been identified? Is it clear how states can make the best possible use of existing data? What are the next steps in this area?

Is it clear how should children with special needs (developmental disabilities, LEP, at-risk) should be included in assessment and data gathering? What are the next steps in this area?

1:00 to 1:30 Lunch

1:30 to 2:30 Developing Comparable Results and Data

How can results be developed and data collected that reflect community/state input and priorities and that are also comparable across communities and states? Should we strive to use the same data collected by communities and states (increasingly for the purpose of higher-stakes accountability such as resource allocation) to make comparisons among communities and states? to generate a picture of children and families across the nation? Alternately, should the goal be separate local, state, and efforts to define results and collect data? How could separate local, state, and national data collection efforts support one another and minimize duplication? What are the next steps in this area?

2:30 to 3:00 Involving the Early Childhood Education Community

How can the early childhood education community be better integrated into efforts to develop and assess child-based results? What are the next steps in this area?

3:00 to 3:30 Public Relations

How can states and communities develop the political attention and will to overcome special and competing interests and come up with fair and meaningful results and assessment systems for children and families? What are the next steps in this area?

3:30 to 4:30 Finalize Next Steps

Second Issues Forum: Next Steps for Child-Based Results
January 24, 1996
Participants

Barbara Blum
Foundation for Child Development

Barbara Bowman
Erikson Institute

Sue Bredekamp
National Association for the Education of Young Children

Cynthia Brown
Council of Chief State School Officers

Nancy Cohen
Bush Center in Child Development and Social Policy,
Yale University

Ellen Galinsky
Families and Work Institute

Eugene Garcia
Graduate School of Education
University of California Berkeley

Sharon L. Kagan
Bush Center in Child Development and Social Policy
Yale University

Luis Laosa
Educational Testing Service

Michael Levine
Carnegie Corporation of New York

John Love
Mathematica Policy Research, Inc.

Kristin Moore
Child Trends

Michelle Neuman
Bush Center in Child Development and Social Policy
Yale University

Gregg Powell
National Head Start Association

Sharon Rosenkoetter
Associated Colleges of Central Kansas

Jack Shonkoff
Heller Graduate School
Brandeis University

Gary J. Stangler
Missouri Department of Social Services

James Ysseldyke
National Center on Educational Results
University of Minnesota

Appendix C: *Considerations Regarding an Results-Based Orientation*

Sheldon H. White
Harvard University

Paper presented at the Issues Forum on Child-Based Results, New York City

SECTION 1 *Introduction*

In the abstract, program assessment using child-based results is highly desirable for the management of early childhood programs. To the extent that a program produces changes in children that are unarguably positive and plainly visible to others, those responsible for the program have less to worry about internally and less to explain to others. The program takes care of itself. The program managers obtain autonomy and flexibility. Supervisors or critics may ask questions about the program's philosophy, methods, or operational strategy—but if the positive benefits of the program are plain to see, all the questions are held at arm's length. They do not disappear, they are tabled, but they have little force in requiring changes in the program. Creative program developers attach considerable importance to the "immunization" produced by face valid results. I have seen one sophisticated developer of an innovative educational program go to considerable effort to produce a new system of evaluation along with his new educational program. His hope was that his new form of evaluation would place a shield between his unorthodox program and its critics.

Often enough, the results obtained by an early childhood program do not immediately and obviously declare themselves as benefits. Then some kind of estimation of what the program is achieving has to be arrived at by proxies: goodness-of-process indicators, peer reviews, surveys of client satisfaction, and analyses of outcome variables that might be theoretically or argumentatively linked to the possibility of future benefits. Judgments about a program using such catch-as-catch-can indicators may differ. Such indicators may be reasonable and adequate for the everyday management of a program, but they may not be sufficient to answer life-or-death challenges to the program in a political context.

The travails of Head Start are, of course, a perfect illustration of the challenges that confront a program supported by partial and argumentative indicators. Child-based results are inherently difficult for many early childhood programs because the important consequences the programs are intended to influence lie far in the future. Either we look at some short-term proxies for those distant consequences, or else we have to wait a long time before finding out whether or not the program has had an effect. When there are life-or-death issues of accountability—as, for example, when questions about Head Start's validity periodically surge forth in the Congress—we struggle with the choices.

SECTION 2 *The State-of-the-Art of Readiness Testing*

Since Head Start is generally understood to be a program that helps disadvantaged children do better in school it would have seem natural to have used tests of school readiness as indicators of the program's effectiveness. Such tests have hardly been used in the great volume and variety of Head Start studies. Why would people not use a test so patently and obviously directed towards just what Head Start is supposed to bring about? The technical qualities of the tests are not very good and this is a reflection of the fact that what the test is intended to deal with is very poorly understood.

To begin with, traditional readiness tests are not very good in conventional psychometric terms. About a decade ago, in conjunction with my longstanding interest in developmental changes in children in the 5-7 age range, I looked at a number of the major commercial school readiness tests. I was interested in the possibility that there might be some hidden wisdom deep in their construction. Did readiness tests embody insights about the cognitive changes in children near the onset of schooling? Did the subtests of those instruments differentiate out theoretically interesting factors in children's cognitive development? I had to be interested in the tests' predictive power. Could the tests do what they were designed for, assess children's cognitive maturity?

The readiness tests did not look as though they contained much hidden wisdom. They looked like work-samples of things that children are asked to do in the early grades. But the readiness tests' statistics said the predictive power of the tests was so poor as to make theoretical questions about the instruments uninteresting. I did not pursue the analysis at that time, but it was the memory of it that

led me to be personally quite skeptical when, a few years ago, I first confronted a declaration of Goal 1 of the Goals 2000 project.

Within the past few weeks, in preparation for today's meeting, I have taken a second look at the readiness and readiness-like tests now on the market. Table 1, which can be found at the end of this paper, gives some statistics on the tests. Fifteen readiness tests were looked at—the Woodcock-Johnson Psycho-Educational Test Battery, the Brigance Diagnostic Inventory of Basic Skills, the Howell Pre-Kindergarten Screening Test, the Brigance K and 1 Screen, the Daberon Screening for School Readiness, the Anton Brenner Developmental Gestalt Test of School Readiness, the Analysis of Readiness Skills, the Metropolitan Readiness Tests, the Clymer-Barrett Readiness Test, the Basic Skills Inventory, the Gesell School Readiness Test, the Cognitive Skills Assessment Battery, the Lollipop Test, the ABC Inventory to Determine Kindergarten and School Readiness, and the McCarthy Screening Test. Table 1 also gives the same information for a second set of eight tests of general intellectual development—the Boehm Test of Basic Concepts, the Bracken Basic Concept Scale, the CIRCUS, the Battelle Developmental Inventory, the DIAL, the WIPPSI, the ABC Inventory, and the EARLY. Table 1 gives each test's declared purpose, the age range for which it is intended, and some statistics on reliability, concurrent validity, and predictive validity.

The statistics for this 1995 sample of readiness tests are, on the whole, slightly better than those I remember from the earlier group. Still, predictive validity information is missing or vague for many of the tests. Where the tests do predict school-age performances of children, they do not predict very far into the future and they predict to psychometric instruments that are themselves of uncertain predictive power. Interestingly, only two of the tests—the Howell Pre-Kindergarten Screening Test and the ABC Inventory—were correlated with teachers' clinical estimations of whether their children were ready for school, with mixed results.

I have no desire to pass off this quick survey of the current readiness tests as a definitive study. It is not. But the quick survey gives a glimpse of the state-of-the-art of our contemporary capacity to build a strong school readiness test, and that survey is not encouraging about the prospects for building a nationwide school readiness screening instrument by the year 2000.

Of course, there are some good reasons for believing that if we can give some serious, sustained, deliberate efforts to building new school readiness assessment we can do better than those traditional instruments. We know more about the possibilities of testing and assessment, and we have recently begun to modify and diversify our century-old technology of testing. We know more about child development than we used to. And we have accumulated some greater understanding about when and how research data is brought into use in the policy process.

SECTION 3 *Recent Changes in our Understanding of Child Development*

The fundamental reason why traditional readiness tests have not worked well is because we have never had a very clear idea what "readiness" is to begin with. The "readiness" issue arose together with practical efforts to achieve "readiness" testing in the late 1920s and early 1930s. Compulsory education laws were being passed in all the American states and children were pouring into the schools. Some children were visibly less ready to do business in the first grade; teachers could see that. But there was no literature in the 1920s, and there is no literature now, to discuss exactly where or how "readiness" might be constituted in a child.

Two other conceptions that are much like "readiness" exist in the child development literature—Binet and Simon's turn-of-the-century conception of "mental age", built into our contemporary practices of mental testing, and Piaget's notion of "cognitive stages" in childhood, built into many of the "developmentally appropriate" preschool curricula of the present. The Binet-Simon "measuring scale of intelligence" was designed to see if children entering school could profit from regular instruction. Binet and Simon arranged a series of tasks and performances to form an age-scale, and from the score a child obtained on their series they computed a "mental age". Although the Binet-Simon testing procedure has been all but buried under a cloud of subsequent psychometric technology and ideological legendary, the test remains at bottom an instrument for deciding how old a child is mentally. Presumably, there is a uniform path of mental development followed by all children. The goal of the test is to find out how far along the Binet-Simon path the child stands.

A not-dissimilar uniformitarian view of a child's cognitive development came to life in the 1960s, through the enormous influence Jean Piaget had on American developmental psychologists at that time. Many American developmental psychologists who thought about designing and evaluating programs for children in developmental terms conceived of that development as it was construed by Piaget's theory. Child development was cognitive development. All children go through a uniform series of stages of cognitive development, and the important difference between one child and another was the question of how far along Piaget's path each child was. Programs for poor children, it was said, should "close the gap". Some American research on Piagetian theory dwelt on what Piaget called "the American question", the question of whether one could or could not accelerate the movement of the growing child along Piaget's path.

Piagetian theory is on the wane now, and we have come

to believe that there is more to the small child's life than marching from one Piagetian stage to another. There is motor development, social and emotional development, the organization of language, the building of general knowledge, and the building of metacognitive insights and strategies. One of the heartening things about the contemporary literature on child development is that all these aspects of child development are being actively studied. I believe we know far more about children's development than we have so far "harvested" for the design of programs and assessment instruments for children. We can make a richer world of programs for children and families with such knowledge.

But we have to proceed with care and with thoughtful and sophisticated efforts towards instrument development. Some of the simplifying assumptions of the past are now slipping away from us—the notion that all there is to child development is cognitive development, and the uniformitarian notion, the idea that all children pursue a common path towards adulthood.

Once upon a time, the heart and soul of early education was the celebration of the diversity of small children. When the Froebelian *Kindergartens* first came over to the United States, the women who worked in them called themselves "child gardeners." Children differ. One is a tomato, another a carrot, a third a sweetpea, a fourth a cucumber. The task of the child-gardener is to study each child and to see the way the way it grows and what it needs, and then to offer that child the support or guidance best suited to his or her needs. The labor-intensive vision of how to adult should deal with her small children in a 19th century Froebelian kindergarten is far behind us now. Our children go to modern kindergartens and grade school classes in which they pursue common schooling—the "basics" of literacy and numeracy we expect will be given to all children. We like uniformitarian visions of the basic processes of child development because they dovetail nicely with the standardization of our classrooms and our expectations. But the reality that every teacher and parent knows is that children are different from one another. If we are going to assess children's readiness in a broad way, we are going to have to address those differences in some meaningful way.

Consider social development, which everyone now agrees is an important aspect of what happens to children in preschools and schools. What if one shape or form of social development that is true for all children does not exist? Children behave differently in social situations. Some are bold, some are shy, some are friendly, some are reserved, some are garrulous, some are taciturn. Most children in preschools and schools manage to solve the problem of finding a comfortable social existence among their peers. They catch hold of some sector of small-fry society, but they do not all do so in the same way. Children are very much like adults in this regard. It is not clear which uniform criteria of social "readiness" for schooling one can apply to all children, boys and girls, who are members of one cultural community. It is even less clear which criteria of social development should be applied to children of different cultural communities in American society.

If we are going to set forth readiness tests for schools that examine a broad spectrum of children's functions and capabilities, I believe we are going to have to confront and deal with the non-standard, idiosyncratic aspects of individual children's development. The community of developmental psychologists now includes strong groups of researchers addressing each of the several major streams of small children's development, and I suspect we can work with such researchers to gradually begin to develop possibilities for broad-scale examinations of children's capabilities and competence in several areas of development. But I see no quick way to carry out this process, and I am quite certain that we cannot smash-and-grab our way past the necessity for entering into it. The research and development processes that will be entailed will extend considerably past the year 2000.

Politicians like to set forth lofty and heroic goals. It is in the nature of leadership that they do so; 'impossible' goals get peoples' attention, mobilize them, and surprisingly often turn out to be possible after all. Furthermore, the life of high officials nowadays tends to be dull, nasty, brutish, and short. One reason why projects on behalf of children tends to be framed in impossibly short periods of time is the brief time officials in power have to act. Understanding all that, I am still not persuaded that we can or should try to establish a nationwide system of readiness assessment by the year 2000. We have had enough 20th-century experience with the development and use of psychoeducational tests to know that testing is a double-edged sword. It can hurt as well as help.

Tests that teachers and administrators do not respect can be more or less politely subverted or evaded. There can be minor forms of fraud, as when 50 American states all report that their children are above average on the school achievement tests—the famous "Lake Woebegon Effect." Tests can control and limit what teachers do, as when teachers in many American classrooms set aside their best professional judgment and "teach to the test". And some psychoeducational tests, notoriously the descendants of the Binet-Simon instrument for determining children's mental age, can become significant instruments for inter-ethnic politics and ideology. Our experience with psychological test usage to date suggests that we have every reason to be slow and careful in our development of future instruments.

SECTION 4 *Recent Changes in Testing*

I am optimistic about the possibilities of more sophisticated assessments of children's readiness, given time. A reasonably restrained process of test development can do much to

move us towards a greater ability to use child-based results for such purposes as program development and evaluation, student assessments, teacher guidance, policy determination, and a variety of other practical uses. I began my talk today with a consideration of the state of the art of commercial tests for assessing children's readiness. In closing, it might be useful to consider again briefly what is happening in the world of commercial testing. The psychoeducational enterprise traditionally known as "tests and measures" is undergoing a great deal of change today, after a good many decades of being surprisingly stable and resistant to change. The more important changes of the present are the following:

Different Kinds of Testing are Being Developed for Different Purposes

Traditional psychological tests have always been a little like Henry Ford's Model T. Ford would sell you any color car you wanted so long as it was black. Traditional psychoeducational tests could be used for any purpose that you wanted as long as you used a standardized, forced-choice, norm-referenced instrument. But it is not at all clear that one kind of test is maximally useful for purposes of federal- or state-level accountability, providing guidance to a teacher in his or her classroom, diagnosis of an individual student's strength or weaknesses, or assessing the pros and cons of a programmatic innovation. Today, we are seeing a differentiation of forms of psychoeducational testing, as differing instruments are being created for different audiences and purposes.

New Technologies of Testing Are Coming Into Use

The differentiation of new forms of testing is being furthered by the emergence of new technologies of testing. The chief modality for psychoeducational testing seems to be still, at the moment, the multiple-choice series of questions addressed with a Number 2 pencil. But a variety of more complex forms of testing are coming forth—ranging from constructed-response testing implemented by paper and pencil or by computer to the evaluation of student work using portfolios or performance observational schemes. In general, the new forms of testing are more expensive to administer and score but they yield much richer and more complex information about the individuals being tested. We can expect that they will play a substantial role in future early childhood programs.

Testing and Teaching Have Begun to Merge

As tests have become more complex, naturalistic, and 'authentic' they have begun to look more and more like extensions of the ordinary learning activities of children in their schoolrooms. Two interesting things have begun to happen. First, teachers have begun to see the tests as interesting commentaries not only upon the child and his or her capabilities, but upon the substance and methods employed by the teacher. The new modalities encourage reflective teaching. And, more and more, it seems likely that they can become part and parcel of the educational process itself. Instead of existing as a diversion or timeout from school work, the new tests become simply an enriched source of feedback coming out of school activities, available to children and teachers who participate in those ongoing activities.

SECTION 5 *Enriching the Mixture of Child-Based Results*

The use of child-based results is not an all-or-none thing in program assessment. We can imagine a future development process in which assessments of programs in early childhood can be progressively enriched through the use of more and more child-based results as we develop the capability to envisage them in a reliable and credible way.

We usually think about accountability in top-down terms, because 20th century discussions of evaluation and accountability have typically arisen within the context of federally managed programs. Higher-order management, providing resources for the early childhood program and answerable for the program in the larger web of government, has to judge what the program is achieving.[1] The program is evaluated, by one means or another. But higher-order management is only one of the parties with an interest in a human services program. Individuals working within the program have an interest, and a real need, to know whether the program is or is not attaining meaningful results. Clients have such an interest; if the program's clients are children, then parents are involved. Professionals and program managers working in the community served by the program have a need to estimate what the program is achieving, in order to come to terms with the possibilities of cooperation or competition with the program.

I believe we should move now to capitalize on the possibilities opened up to us our studies of child development, by the opening up of new forms of testing and assessment, and by our growing understanding of where and how information is used in program guidance and policy formation. We can do this—if we are able and willing to submit to slow, reflective, collaborative processes of instrument development.

1. To simplify discussion, I am here assuming that a government program has one and only one purpose, well-understood by all parties and agreed to by them. But this is not true for a good many government programs that, in the political process, are put forward by coalitions of parties who are directed towards a variety of purposes. Head Start, for example, is generally understood to be a child development program. Historically, however, Head Start was put together by parties with declared major or minor interests in: (a) Civil Rights; (b) community action; (c) the coordination of services for children; and (d) stimulating school reform. Many programs in the human services reflect such coalitions of interests and emphases.

TABLE 1

TECHNICAL ASPECTS OF SCHOOL READINESS TESTS

Test	Age	Test-Retest	Inter-rater	Reliability	Concurrent Validity	Predictive Validity	References
Lollipop Test (1981)	Pre-K to 1st			KR20 for whole test = .90	.58 w/ teacher ratings .86 w/ MRT	Lollipop given at end of K; MRT after 1st = .73; MRT after 4th = .40	
Bringance K and 1 Screen (1986)	K to 1st						
Bohem Test of Basic Concepts, R. (1986)	K to 2nd	K-.88 1st-.55 2nd-.66		.62 to .82	.60 w/ PPVT .24 to .64 w/ Comprehensive Test of Basic Skills, California Achievement Test, Iowa Test of Basic Skills		
Bracken Basic Concept Scale (1984)	2-6 to 7-11	.97 for total test		.76 to .80	.68 to .88 w/ PPVT, Bohem, MRT, and Token Test		
McCarthy Screening Test (1978)	2-6 to 8-6	.32 to .69		.41 to .80	.66 w/ Peabody Individual Achievement Test		

Test	Age	Test-Retest	Inter-rater	Reliability	Concurrent Validity	Predictive Validity	References
Battelle Developmental Inventory (1984)	0-6 to 8-0	> .90	> .90	.81 to .95	.41 to .60 w/ Stanford Binet		T.C. II, 72-82
Gesell School Readiness Test	4-6 to 9	.79	.87	.84	83% w/ teacher ratings; .64 w/ Piagetian test battery; .50 w/ Thorndike IQ's; .61 w/ Thorndike Mental Ages	Gesell in K; .64 w/ Stanford A.T. in 1st	
Metropolitan Readiness Tests (1986)	PreK to 1st	.62 to .92		.66 to .93		Took MRT; 6 months later, .34 to .65 w/ Metropolitan Achievement Test; .47-.83 w/ Stanford A.T.	
Wechsler Preschool and Primary Scale of Intelligence (1967)	4 to 6-6	.86 to .92		.77 to .96	.75 w/ Stanford Binet .58 w/PPVT .64 w/ Picorial Test of Intelligence		
Wechsler Revised (1989)	2-4 to 7-3	.81		.91 to .96	.74 to .90 w/ Wisc-R, Stanford Binet, and McCarthy		

Test	Age	Test-Retest	Inter-rater	Reliability	Concurrent Validity	Predictive Validity	References
Developmental Indicators for the Assessment of Learning, R. (DIAL) (1984)	2 to 6	.76 to .90		.96	.40 w/ Stanford Binet		
Basic School Skills Inventory (1983)	4-0 to 7-5			.88 to .92	.22 to .43 w/ teacher ratings		T.C. IV, 68-75
Chicago Early Assessment and Remediation Laboratory (EARLY) (1984)	3-0 to 6-0	.72 to .91	.89				ERIC ED 204 372
CIRCUS (1979)	PreK to 1st			.74 to .89			T.C. VII, 102-109
Cognitive Skills Assessment Battery	PreK to K			.80			T.C. VII, 126-139
Developing Skills Checklist (1990)	PreK to K; 4 to 6						N/A Yet

Appendix D: *Judging Interventions by Their Results* *

Lisbeth B. Schorr
Harvard University Working Group on Early Life

Paper presented at the
Issues Forum on Child-Based Results
New York City

Section 1: Introduction

I have been asked to discuss how child-based results in early childhood are related to the broader political context in which the shift to results accountability is currently occurring. That political context is defined, in my view, first, by an increasing concern about the state of America's children and families, particularly about escalating rates of violence among ever younger children, of children bearing children, and of youngsters coming of age without the skills or motivation to earn a decent living. Secondly, it is defined by a growing sense that nothing systematic can be done—except, perhaps, for harsh punitive measures—to reverse these trends.

I believe that a shift toward results accountability is an essential strategy in efforts to build on new research and experience to improve results for children and families. Results accountability is not a panacea, but it could be a major step toward improving the conditions in which children grow into adulthood. It could also become a central strategy in a concerted effort to improve results for children growing up in high-risk environments.

BACKGROUND

While much of the current discussion about whether anything works, whether government does anything right, and whether the government that governs least also governs best, is pure rhetoric, the rhetoric hides some real issues that need urgently to be addressed. Among them is how to improve our ability to differentiate what works from what does not. Legislators have to know what works when voting on laws and appropriations, parents want to know whether their child's school is providing an effective education, foundations have to know whether they are supporting a promising strategy, and voters who have given up on compassion want to know what is a good investment.

Until recently, anyone who wanted to know whether tax or philanthropic dollars were being spent for a good purpose was offered one of three unsatisfactory responses: The most traditional response has been to bypass the problems of obtaining information about results and to assume that what mattered were intentions and efforts, institutions and services, resources and spending (Manno, 1994). The family service agency was doing its job if its budget was increasing and its monthly parent education sessions were attended by a specified number of people and were under the supervision of a certified social worker.

A second, more recent, response has been to say that if you really want to know what is working, you have to privatize the function—let the market place become the judge of effectiveness, by shifting the school or day care or recreation or mental health program out of the public or nonprofit sectors. "Shift the burden of evaluation from the shoulders of professional evaluators to the shoulders of clients, and let them vote with their feet," advises UCLA professor James Q. Wilson (cited in Dilulio, 1994, p. 58)[1]. And that is how, presumably, we know preschool programs work—middle class parents spend money on them.

In a third response, providers of health, education, and social services say "Trust us. What we do is so complex, so hard to document, so hard to judge, and so valuable, in addition to which we are so well intentioned, that you, the public, should support us and our programs without asking for evidence of effectiveness. Don't let the bean counters who record the cost of everything and know the value of nothing interfere with our valiant efforts to get the world's work done."

Since the mid-1980s, in the face of ever-increasing skepticism about the value of public investments in any human services, a fourth answer has emerged. A new breed of social reformer is contending that public support for social investments would be greatly strengthened if citizens, tax payers, customers, clients and communities were able to hold the providers of services, supports, and education accountable for achieving the results that citizens value. Many reformers are also coming to see results-based accountability as an important way of increasing program effectiveness by freeing human services from the straightjackets of rigid rules.

Section 3: The Problems That Results-Based Accountability Can Solve

GIVE THE PUBLIC SOME PROOF OF RESULTS

Large numbers of US citizens have a deep sense that they are not getting their money's worth from their governments.[2] The 1995 confirmation hearings on the nomination of Dr. Henry Foster to become Surgeon General featured lengthy

and often confused exchanges on the impact of "I Have a Future," the teenage pregnancy prevention program which Dr. Foster founded in Nashville, Tennessee. After much discussion about the meaning of several program evaluations, Senator James Jeffords of Vermont finally stated, in some exasperation, "We're fooling ourselves to think these programs are good because they feel good, when the evidence of impact isn't there." Senator Jeffords is not alone in his sense of frustration. In fact he speaks for an increasing number of government officials and citizens whose faith in social programs will not be restored until they know what, exactly, they are getting for their money, be it tax money or large-scale philanthropy.

Paying attention to results rather than inputs is central to the reinventing government proposals of David Osborne and Vice President Al Gore. But they were hardly the first to preach the results gospel.[5] Two decades before she became head of the Office of Management and Budget in the Clinton Administration, Alice Rivlin was calling for better measures to assess the success of social programs, because public concern with ineffectiveness of human services was running "very high indeed." (Rivlin, 1971, p. 65). She concluded that "all the likely scenarios for improving the effectiveness of education, health, and other social services dramatize the need for better (outcome) measures. No matter who makes the decisions, effective functioning of the system depends on measures of achievement. .... To do better, we must have a way of distinguishing better from worse" (Rivlin, 1971, pp. 140-41, 144).[6]

FREE HUMAN SERVICE PROGRAMS FROM THE STRAIGHT-JACKETS OF CENTRALIZED MICROMANAGEMENT AND RIGID REGULATION

Management by results is the best alternative to the top-down, centralized micromanagement that holds people responsible for adhering to rules that are so detailed that they make it impossible for a program or institution to respond to a wide range of urgent needs.

Whereas the bureaucratic paradigm assumes that control can only be exercised by rules, an outcome oriented organization substitutes "adherence to norms" to fulfill the same function (Barzelay, 1992, pp. 124-5). A commitment to results is essential to this shift, because a clear understanding about purposes and desired results is the basis on which employees will take responsibility for adhering to norms, and will channel their energies into making appropriate adaptations and solving problems. Employee performance improves when employees feel accountable because they believe that their intended work results are consequential for other people (Barzelay, 1992). An results orientation also encourages staff to think less categorically as they become more aware of the connection between what they do and the results they seek.

ENHANCE SOCIETAL AND COMMUNITY CAPACITY TO BE MORE PLANFUL AND MINIMIZE INVESTMENT IN ACTIVITIES THAT DO NOT CONTRIBUTE TO IMPROVED RESULTS

Agreement on a common set of goals and outcome measures makes collaboration easier, and also fuels the momentum for change and helps promote a community-wide "culture of responsibility" for children and families.

Reflecting Alice in Wonderland's insight that if you do not know where you are going, any road will get you there, a focus on results is likely to discourage expenditures of energy, political capital and funds on empty organizational changes and on ineffective services. The shared commitment to improve results for children is what can make efforts at collaboration and service integration fall into place—not as an end, but as an *essential means* of working together toward improved results.

FOCUS ATTENTION ON WHETHER INVESTMENTS ARE ADEQUATE TO ACHIEVE THE PROJECTED RESULTS

The new conversation about results may have its most profound effect by injecting a strengthened ethical core into human service systems' that currently focus more attention on the fate of agencies and programs than on whether people are actually being helped. The new results focus promises (or threatens, in the eyes of some) to end a conspiracy of silence between funders and program people by exposing the sham in which human service providers, educators, and community organizations are consistently asked to accomplish massive tasks with inadequate resources and inadequate tools. Attention to results forces the question of whether outcome expectations must be scaled down, or interventions and investments scaled up to achieve their intended purpose.

In the past, parent education programs have been funded with the vague expectation that they would somehow reduce the incidence of child abuse, although a few didactic classes have never been shown to change parenting practices among parents at risk of child abuse. Similarly, outreach programs to get pregnant women into prenatal care are expected to reduce the incidence of low birth weight, based on the similarly vague belief that outreach programs are a good thing, without any knowledge of whether the prenatal care that is made more accessible actually provides the services that could be expected to result in a greater number of healthy births.

Especially in circumstances where it will take a critical mass of high quality, comprehensive, intensive, interactive interventions to change results, where effective interventions must be able to impact even widespread despair, hopelessness and social isolation, funders and program peo-

ple should resist the temptation to obscure the limitations of so many current efforts. Providers—and even reformers—who are asked to achieve grand results with interventions so paltry that they are in no way commensurate to the task, should not obscure the insufficiency of the investment by pleading with funders and evaluators to just document their efforts and not their results because it would not be fair to hold them accountable for real results changes when they are doing the best they can. Evidence that a diluted form of a previously successful intervention is not making an impact is not an argument against results-based accountability. It helps to clarify that dilution regularly transforms effective model efforts into ineffective replications. Recognition that a single circumscribed intervention may not be sufficient to change results is not an argument against results-based accountability. It is an argument for adequate funding of a combined critical mass of promising interventions.

SECTION 3: RESULTS-BASED ACCOUNTABILITY: A FAUSTIAN BARGAIN?

Critics of the push toward results accountability range from the skeptical to the appalled. Commenting on pressures to incorporate results accountability in early childhood programs, Sue Bredekamp, of the National Association for the Education of Young Children, says it is but "one more opportunity and justification to 'blame the victims,' because the children who are in greatest need of services demonstrate the poorest results" (Bredekamp, 1995).

Skeptics see the willingness to be held accountable for achieving specified results as a Faustian bargain—even when they agree that a shift toward results accountability has the potential to solve a lot of serious problems. They believe that human service providers, in their eagerness to obtain more funding and to escape over-regulation, will become unwitting tools in the war against government, the war against the vulnerable, and the war against all public sector activities that are grounded in considerations of morality, ethics, and social justice.

FEARS OF RESULTS ACCOUNTABILITY

Those who resist the push to results accountability have at least six specific fears:

First, knowing that what gets measured gets done, they fear that programs will be distorted. What will get done will be what is easiest to measure and has the most rapid payoff—rather than what is really important. They point out that most communities have aspirations for their children that greatly exceed the results that are currently measurable, especially when the demand is for quick evidence of success. Will community health clinics raise immunization rates at the cost of cutting back on other kinds of well child care, or support for chronically ill children? Will preschool programs deprive children of the opportunities for play that stimulates creativity and teaches empathy in order to reserve more time for the flash cards whose mastery shows up on "school readiness" tests?

A second fear is that even effective programs will seem to be accomplishing less than they actually are. If an inner city consortium gets funding to improve the employability of youngsters coming out of high school, will it be judged a failure if the predictions that produced the funding turn out to be overly optimistic? And will it be judged a failure if results do not change as quickly as the funders had hoped? Will the consortium be held responsible for achieving citywide improvements in results even though they were only able to work with only 150 youngsters and their families?

A third fear arises from the recognition of the complexity of the most promising interventions, and the corollary that in the complex, interactive strategies that are most promising, responsibility for both progress and failure cannot be accurately ascribed. No single agency, acting alone, can achieve most of the significant results. If higher rates of children ready for school depend on the effective contributions of the health system, family support centers, high quality child care, nutrition programs and Head Start, as well as on informal supports and community activities, will agency accountability be weakened as attention shifts to communitywide accountability efforts? How are agencies to be held accountable for results over which no single agency has control?

A fifth fear is that results accountability will become a shield behind which the few remaining protections and supports for vulnerable children, youth, and families will be destroyed. Especially in this anti-regulation era, rock-bottom safeguards against fraud, abuse, poor services, and discrimination based on race, gender, disability, or ethnic background could be destroyed. The new results orientation could lead to the abandonment of the input and process regulations that now restrict the arbitrary exercise of front-line discretion by powerful institutions against the interests of powerless clients.

A sixth fear is that the results-based accountability that is intended initially to serve such benign functions as creating pressures for reform and sharpening the focus of managers and practitioners on accomplishing their mission rather than preserving their turf, will soon lead to such hard-edged consequences as results based budgeting. The actual allocation of funding based on a program's or agency's performance could ultimately threaten to shut down programs and agencies that cannot provide evidence of their contribution to achieving agreed-upon results, despite the fact they may be making such a contribution.

Section 4: The Special Case of School Readiness

The first of the national education goals—agreed to initially by the nation's governors and President George Bush, and subsequently endorsed by both President Bill Clinton and the U.S. Congress—was that by the year 2000 all children will start school ready to learn. The importance of this goal is not disputed. Recent research has it made quite clear that brain development occurs earlier, more rapidly, and is more vulnerable to environmental influence and more lasting than had been previously been suspected (Carnegie Task Force on Meeting the Needs of Young Children, 1994).[7] Young children's experiences between birth and school lay the foundations for success in school and in life (National Research Council, 1991). There is evidence that children without a good foundation *do not* do well at the beginning of school and will not be able to catch up. In addition, the proportion of children in a kindergarten or first grade class who are ready for school learning is a powerful predictor of how much learning will go on in that class.[8]

Despite the unanimity around the importance of school readiness, the prospect of measuring it has aroused deep passions and controversy. This may be because the welcome recognition among policy makers and the public that the early childhood years are such a critical time has led to decidedly unwelcome consequences: the use of standardized tests to make high stakes decisions about individual children, including labeling some unready for school entry and placing others in special tracks. The result has been that a high proportion of children have a damaging "failure" experience at the very beginning of their academic careers, and that some preschool programs have been driven to teaching test-taking skills, concentrating on rote learning, memorization, and drill ... rather than on the exploration and experimentation and grasp of basic concepts that are key to later learning .... and that foster confidence, curiosity, and problem solving" (National Research Council, 1991, p. 2, 10).

These developments have left the early childhood community so traumatized at the possibility of unwittingly promoting further inappropriate testing, that many oppose any attempt to assess school readiness by testing or observing individual children, even if the testing is done for the purpose of judging the community's provisions for preparing children for school entry, and not the abilities or capacities of individual children.

There is little dispute about the elements of early experience that contribute to school readiness. They include good health care and nutrition, high quality child care and preschool experiences, communities that support families, and homes that provide children with the conditions that develop trust, curiosity, self-regulation, the foundations of literacy and numeracy, and social competence (Action Team on School Readiness, 1992; Boyer, 1991; National Task Force on School Readiness, 1991; Zill, 1995). There is a school of thought that concludes that we know enough about what communities have to do to produce these results, that it is possible to measure community capacity to assure school readiness as a proxy for measuring children's school readiness directly. The National Governors' Association, for example, has proposed a list of community capacity indicators for this purpose.

My own view is that community capacity indicators would serve as a reliable proxy if we knew more than we now do about the precise linkages between inputs and results, between, say, home visiting and infant health or between family support centers and parental competence, between the elements of good child care and the development of curiosity in toddlers.

However, given the current state of knowledge, measures of community capacity will be informative but not definitive in assessing progress toward universal school readiness. While it is absolutely clear that the extent of school readiness depends on the existence, accessibility and quality of an array of services, supports, and institutions, it can probably best be discerned by looking directly at samples of children.

The question then becomes, can children's school readiness be determined without doing them any harm? Can it be done in ways that would make it impossible to label or stigmatize individual children? Can it be done in ways that would strengthen rather than distort preschool programs that "taught to the test"? Can it be done in ways that "acknowledge the fluid and cumulative nature of development" and that do not result in "blaming children and families for low levels of early learning?" (Phillips & Love, 1994). Can it be done in ways that make clear that if large numbers of children are not ready for school, that is not a child problem but a community problem?

A great deal of work, such as that led by John Love at Mathematica Policy Research, has been done to suggest that these questions can now be answered affirmatively (Love, 1995).[9] But a clear consensus in the early childhood community has not yet emerged around this proposition. It may be that assessments of school readiness in the most immediate future will have to rely on approaches that combine observations of samples of children and measures of community capacity, as proposed by Yale Professor Sharon Lynn Kagan in the most recent National Governors' Association Issues Brief (Kagan, 1995).

Section 5: Choosing the Right Results

When communities or states (or the nation, in the case of the education goals) actually agree on results that all the stakeholders consider important and meaningful, a lot of

other things fall into place. Results that have been agreed upon by professionals and clients and other interested parties can become a solid foundation on which new strategies can be hammered out, and flexible responses can be adapted and evolved on the basis of continuing feedback." With results as constants, we can afford to experiment and even disagree over the means: whether and for whom home visiting is more effective than family support centers; whether parent involvement is best achieved by helping parents to read to their children or help them with their school work, through parent employment as classroom aides, or through parent participation in governance; whether children best learn to read using phonics, whole language, or some other method. By being moored to the ends, it is possible to stay flexible on the means.

MEASURING SUCCESS

Of course, if results are to be the constants, the selection of outcome measures becomes enormously important. For better or worse, what gets measured has a great effect on what gets done. In one way or another, teachers teach to the test, just as social workers pay attention to what the auditors count. So the trick is to devise measures that come as close as possible to actually reflecting what ought to get done. If you want children to learn to reason in math, you go beyond multiplication tables in assessing their performance. If you want children whose eyesight is defective to be treated, you measure the absence of untreated vision defects among first graders, not the number of vision screenings that have been done.

In addition to all the technical considerations that determine the choice of outcome measures (Love, 1995; Moore, 1994), the greatest stumbling block for those actually engaged in moving to outcome accountability has been the difficulty of forging agreement on a set of results considered important, meaningful, and measurable by a wide range of stakeholders, including skeptics. This involves the resolution of two major tensions: (a) between the implicit and explicit purposes of the efforts, and (b) between all that the community wants from the effort versus what can be agreed upon and measured.

We have been more opportunistic than planful in this country about measuring success. We grasp for what is easiest to measure rather than what best reflects a program's purpose. Thus the Westinghouse Learning Corporation, under a government contract to assess the effects of the earliest years of Head Start, measured only changes in the IQ of participating children, even though IQ is one of the least malleable of human characteristics, and despite the fact that Head Start aimed to improve the health and nutrition and social skills of participating children, to empower their families, strengthen their communities, and in many other ways to contribute to their school readiness. Some thirty years later, IQ still becomes the outcome that gets assessed because it's such a handy measure. Even today family support programs are evaluated on the basis of their effect on the IQ of participating children.

Hard-edged thinking about purposes and results means asking anew about what is worth doing, and to what end. The results around which it seems to be easiest to get agreement are those around which judgments are seen as least subjective, and where, most people agree on the desirable direction of change," even if people do not agree on what they would give up to achieve such change, how to achieve it, or which results are most important (Rivlin, 1971).

But it gets harder, once we get beyond a few outcome measures that are readily agreed upon. It means becoming explicit about everything, including multiple or even conflicting goals. The debate has to include the question of whether the provision of jobs for local residents is among the purposes of Head Start, inner city hospitals, schools, child care or construction projects." Hard-edged thinking about results means acknowledging that while full-day Head Start is likely to pay off in higher rates of school readiness, part of its impact lies in its ability to provide jobs to neighborhood residents, the decreased social isolation of Head Start parents, and the increase in the number of adults who are confident that their children are well taken care of and are therefore better able to pursue job training or employment.

Efforts to define results must also resolve the tensions between all that the community wants from the initiative as against what can be agreed upon and measured. These tensions are especially troublesome around results in early childhood and adolescence.

Many leaders in early education, whose hearts and souls are committed to celebrating the diversity of small children (White, 1995), are convinced that whatever outcome measures are selected to document school readiness, they will "mislabel miscategorize, and stigmatize children" by measuring only narrow cognitive development, (Kagan, 1995) and will result in narrow, standardized, cognitively oriented preschool programming.

Having observed many recent efforts to select outcome measures in a wide variety of contexts, I have become convinced that these tensions will be resolved through increasing recognition of three fundamental points:

> Communities—and certainly parents—have *goals* for their children that are more ambitious, more differentiated, and more nuanced than the results that can be agreed upon and measured

> It is possible to maintain a strong commitment to a *programmatic* orientation that is ambitious, differentiated, and nuanced, while being held accountable to the accomplishment of more modest results

Results selected for accountability purposes must be *persuasive to skeptics*, not just to partisans of the programs and policies being assessed

AMBITIOUS GOALS AND MEASURABLE RESULTS

When communities, parents, practitioners, policy makers and advocates are asked about their goals and their vision for their children, they talk about wanting all children to grow up in loving, nurturing and protective families, to be connected to those around them, and to achieve their personal, social, and vocational potential. They talk about wanting youngsters to feel safe, to have a sense of self-worth, a sense of mastery, a sense of belonging, a sense of personal efficacy, to be socially, academically, and culturally competent, and to have the skills needed for productive employments

Such goals can become a framework within which outcome measures can be selected for accountability purposes, with the understanding that only some aspects of these goals can currently be measured with widely available data and with outcome measures around which it is possible to gain widespread agreement. There is a direct connection between these goals and the outcome measures used for accountability purposes, but goals and outcome measures serve different purposes. Goals represent what the community is striving for. Outcome measures represent what the community will be held accountable for—by public and private funders and perhaps by higher levels of government. The goals can be general, but the outcome measures must be so specific, the public stake in their attainment so clear, and their validity and reliability so well established, that the community would ultimately be willing to see rewards and penalties, as well as resource allocation decisions, attached to their achievement.

A commitment to more visionary goals is entirely compatible with a commitment to documenting progress toward the achievement of these goals by the use of more modest outcome measures. Of course health is more than the absence of disease, educational attainment is more than not dropping out, nurturing family life is more than the absence of abuse and neglect, economic well-being is more than living above the poverty line, and a thriving community is more than an absence of boarded up houses and open drug markets (Brandon, 1992).* The attainment of modest but measurable results would signify substantial progress toward more ambitious goals.

AMBITIOUS PROGRAMMATIC FOCUS AND MEASURABLE RESULTS

Accountability systems that rely on results that are easily measured and are persuasive in a public policy context do not preclude a much broader programmatic agenda. Strategies to achieve measurable results must of course focus on child's play as part of child care, on social skills as well as cognitive skills among preschoolers, on youth opportunities as part of youth development, and on caring and connectedness as part of community building.

RESULTS THAT ARE PERSUASIVE TO SKEPTICS—AND THE SAGA OF EDUCATION STANDARDS

The most compelling lesson about the importance of selecting results that are persuasive to skeptics comes out of the recent experience with national efforts to agree on education standards.

In elementary and secondary education, the shift in judging quality from the amount of per pupil spending, to an approach that asks what children are learning has been occurring rapidly and tumultuously. Outcome standards for student learning were blessed in 1989 by the nation's governors at the Williamsburg education summit called by President Bush, and were embraced by many educators, reformers, and advocates as a way of advancing the twin goals of educational excellence and equity (Manno, 1994).

Despite the fact that there was little agreement on how student achievement should be assessed, the Educational Commission of the States reported that by 1993 states had developed or implemented an outcome-based approach to education (Manno, 1994). This relatively smooth progression did not last long. By February of 1995, Chester Finn, an early promoter of national education standards, entertained a Brookings Institution conference with a talk entitled "The, short unhappy life of national standards." He declared national education standards dead for the foreseeable future. Whether or not his prediction turns out to be correct, the process by which education standards were so quickly and perhaps fatally wounded is one from which those contemplating results-oriented accountability in other arenas have much to learn.

The most fundamental strategic error on the part of the proponents of results standards resulted, in my view, from overreaching. They failed to make the distinction between what they wanted for their children and what they wanted schools, teachers, and their children to be held accountable for.

Proposed draft standards included such items as "All students understand and appreciate their worth as unique and capable individuals and exhibit self-esteem; all students act through a desire to succeed rather than a fear of failure while recognizing that failure is a part of everyone's experiences" (Manno, 1994). From the perspective of a parent, these were reasonable objectives. But whether these should be the objectives of schools or other public institutions, whether they could be agreed on and measured was another question.

Whether the recent push to introduce results standards in education will end up being an impetus or an impedi-

ment to reform is not yet known. What is known is that in the results selection process, it is lethal to ignore the distinctions between the goals that communities have for their children and the results that can be agreed upon and measured, between a programmatic orientation that is ambitious, differentiated, and nuanced and accountability for more modest results, and between results that are persuasive only to an initiative's supporters and those that are also persuasive to skeptics.

Section 6: Mismatch Between the Data That is Needed and the Data Being Collected

There is a severe—and at first blush, strange—paucity of data that could help answer urgent questions about how well efforts to change vulnerable lives are actually helping, and about which efforts help a little, which help a lot, and which help not at all.

The large gap between the data that is needed for results accountability and the indicator data currently being collected exists because data collection has been shaped primarily by only two kinds of pressures: those that reflect the need for administrative data for use in managing programs, policies and institutions; and those that reflect the interests of social scientists, which have focused either on simple indicators that can be monitored for national level trends , or on complex measures of individual development requiring labor-intensive direct observation and data collection. Attempts by policy makers and advocates for children to make do with what has been made available for these other purposes are increasingly unsatisfactory for purposes of designing interventions, holding intervention efforts accountable, and trying to understand what works to change results.

Data collection that has not been required for administrative and managerial reasons has reflected the interests of social scientists who tend to think primarily about national trends, economic influences, and naturally occurring social change. As a result, the nation's data tool kit is virtually useless when it comes to efforts to understand the effects of intentional interventions on the well-being of children and families, especially when those interventions are designed to operate at the level of the local community. The data needed by those who must make judgments about what's working, and who are committed to try to influence current policy debates, is very meager.

Much new work is now needed, and some is beginning to be done,[*] to expand information about the characteristics of children, families and communities that can be reliably measured, and to make data about results for children and families available in a more timely way and in units that correspond to areas that are optimal targets of intervention, such as neighborhoods and school catchment areas. Communities and funders (public and private) need to be able to identify long-term and interim outcome measures that can be linked to interventions, and that can help them assess the effects of interventions across, not just within, the domains of health, education, child welfare, juvenile justice, community development, economic development, and job training. These information needs become increasingly urgent with the need to assess the effects of such new federal initiatives as the empowerment zones and enterprise communities, foundation-funded comprehensive community initiatives for children and families, and impending changes in the allocation of responsibility between the federal government and the states, and in major reforms of such safety-net programs as AFDC, Medicaid, and Food Stamps.

WHO DECIDES?

In determining who selects the results to be achieved, there is much controversy about "top-down" versus "bottom-up" processes. On the one hand, many believe that society has so much at stake in the achievement of a core set of results, that political bodies—probably at the state level—should be responsible for identifying a set out results that are to be achieved in a particular jurisdiction. Others believe that "outcome measures imposed from outside a community have no legitimacy in terms of a local consensus-building process ... and cannot mobilize the resources needed to achieve the results sought ..." (Young, Gardner, & Coley, 1993). Charles Bruner of the Iowa Child and Family Policy Center, who has struggled with this issue both as a state legislator and as an advisor to local programs, argues that those charged with achieving results must be involved in the results selection process if it is to be regarded as fair, useful, legitimate, and if it is to reflect real-life experiences.

There does seem to be increasing agreement that the process of selecting results for accountability purposes, whether it is done by a state legislature or a local collaborative, must have political legitimation. Sid Gardner, of the Center for Collaboration for Children at the University of California in Fullerton, believes that the importance of going through a consensus building process cannot be underestimated, because the selection of outcome measures is not primarily a technical, but a political problem.

It is clear that all of those affected by results-based accountability—as legislators representing tax payers, as providers, or as service beneficiaries or participants—must have a role. All concerned will be able to work more effectively toward common goals if they are able to engage in a consensus-building process, involving both providers and recipients of services, to select the outcome measures they will use or be held accountable by.

Many forms of interactive consultation are possible. For example, when an official state body selects the results, localities may decide or negotiate the numerical value that will represent progress in the achievement of each outcome (e.g., the rate of low birthweight will be reduced by 5% each year, or racial disparities in low birthweight rates will be reduced by 10% each year). But if results are to be used for accountability purposes, and actually carry what David Hornbeck calls hard edged consequences, it seems reasonable that after extensive participation and consultation, the final decisions are made by bodies at a higher or broader level of governance than those being held accountable.

Section 7: *The Importance of Interim Milestones*

The greatest single obstacle to realizing the benefits from a shift to results based accountability is the lack of interim milestones that could reliably show that reform efforts are on track toward achieving their targets.

Local communities, agencies and programs that are struggling to reform, improve, or expand their services, to integrate services across helping systems, or to target a wide array of intensive interventions on selected geographic areas, are clamoring for ways of finding out in the near term whether their efforts are changing results, or even whether they are going in the right direction and making progress toward long-term results. Funders (public and private), practitioners, managers, and systems reformers are becoming increasingly aware that the most frequently cited lesson from major current reform efforts is that they take so much more time than expected—both to get the initiative under way, and to get it to the point where it begins to show an impact on real-world results.[1] They desperately need new tools that would allow them to demonstrate their short-term achievements. They need to be able to get interim information *very quickly*—often long before a program is "proud," (Campbell, 1987) long before it has had a chance to make an impact on rates of school readiness, child abuse, teenage pregnancy, violence, school success, and employment.

Two kinds of interim measures can predict later results: indicators that attach to children, families, and communities and that are a short-term manifestation of long-term results, and indicators of a community's capacity to achieve the identified long-term results.

Examples of interim measures that are a short-term manifestation of long term results for individuals, families, and communities include the following:

Receipt of prompt high quality prenatal care is thought to raise the chances of a healthy birth.

Children who do not read when they are seven are likely to encounter later troubles at school.

An improvement in school attendance rates is thought to predict an improvement in school achievement rates.

Parents' sense of mastery and social support, and the absence of parental substance abuse, as are thought to predict long-term non-recurrence of abuse or neglect.

Knowledge about the connections between measurable indicators of community capacity and long-term results is at a more primitive stage than knowledge about the connections between interim and long term indicators for children and families. Reliable theories about the linkages between interventions and results, and about the constellation of conditions and interventions that will lead to good results, are scarce. Most are unproven. For example, can a community that is developing strategies to reduce rates of low weight births assume with confidence that the "enabling conditions" to reach that outcome are some combination of the capacity (1) to provide family planning services to all persons of child-bearing age, and (2) to provide high quality, responsive prenatal care, nutrition services, and family support to pregnant women?

It is probably not enough to know of the simple existence of certain services, because their quality and how they are made available must be taken into account to link them strongly with results. The distinction among service *availability, access,* and the *nature and quality* of the service in accounting for improved results is crucial—and requires greater understanding and a wider consensus around how to measure the factors that make services effective than now exists (see Charles Bruner's pioneering work).

One connection that most observers consider reasonably well established comes from the early childhood field: a community that is able to offer all of its low income children and their families Head Start and other high quality comprehensive preschool programs is likely to have a high proportion of children prepared for school learning at the time of school entry.

Perhaps the most tantalizing of recently hypothesized links between interventions and results that could produce some new short-term indicators of community capacity are between results for children and families and such indicators of community-level change as a strengthened infrastructure of informal supports, and investments in neighborhood safety and expanded economic opportunity. But there is as yet scant agreement on ways to measure community building, and only modest understanding of the precise connections.

The need for both kinds of short-term indicators that could show movement toward long-term results has long been recognized. It has not been met because the ability to define these interim markers with confidence depends on

BEST COPY AVAILABLE

having reliable evidence, theories, or at least sturdy hypotheses, about the antecedents of major long-term results. Neither social science researchers nor the evaluation industry have really invested in this arena—perhaps because their energies are exhausted by their pursuit of that elusive goal of seeming as scientific as their colleagues in the physical sciences, and because progress in this arena involves a higher ratio of judgment to certainty than most social scientists are comfortable with.

As a society, we now need desperately to make up for lost time. One useful next step would be to systematically examine findings in the recent literature and ongoing experience to provide a more rigorous and deeper understanding of established connections among short-term and long-term results. We need to explore the connections between long-term results on the one hand, and measures of interim individual results and community capacity on the other.[11]

Section 8: *Rigorous Thinking About Process Measures*

Process measures describe what is going on. (Process measures will also continue to be important in assuring that procedural protections are maintained to guard against fraud, corruption, and inequities or discrimination based on race, gender, disability, or ethnic background.)[12] Process measures are an essential component of understanding the impact of an intervention, though they themselves do not assess impact, unless they qualify as interim indicators. Process measures are important in finding ou' whether a program or intervention has actually been implemented according to plan. (Is the Head Start program in operation for the number of hours its funders expect, has it enrolled the expected number of children and the expected proportion of eligible children, has it involved a stipulated percentage of parents, etc.)

Process measures become easily confused with outcome measures and interim measures. Distinguishing among these various indicators is essential to clear thinking about interventions and their consequences. One of the reasons for the current confusion is that the same measure can be an outcome measure, an interim measure, or a process measure, depending on context. If the purpose of the initiative is community building, "community engagement" could be an outcome. If the purpose of the initiative is school readiness and the connection between community engagement and school readiness were reasonably well established, "community engagement" could be an interim measure. If the funder and grantee were to agree to make community engagement an essential component of the intervention, it could be a process measure.

A process measure can be used as an interim indicator if there is a reasonable hypothesis to make the link, as when a program trains parents as community leaders as part of its efforts to rebuild a community infrastructure. A process measure cannot be used as an interim indicator if there is no basis for linking it to long-term results.

The failure to think clearly about process measures, and how they relate to what is being proposed or being done and what is being accomplished, results in what David Osborne calls "process creep" (Osborne & Gaebler, 1993, p. 350). When process creep occurs, means and ends become confused, and the focus on what actually happens to people as a result of the activity is lost. The formation of a collaborative, or a high degree of participation in a new governance entity may be the product of a great deal of effort, but is not, evidence of progress toward agreed upon results unless the rationale that connects these activities to established results is at least explicitly hypothesized, if not proven. The number of children who have been screened for hearing and vision problems is a process indicator. Because screening that isn't followed up with diagnosis and treatment where needed won't reduce the number of children whose vision or hearing is impaired, screening should not be used as an outcome indicator.

But the confusion about process measures is not only conceptual, it is also political. The temptation is ever-present to fall back on using process measures as evidence of progress, even when they meet none of the criteria for outcome measures and there is no basis for linking them to ultimate results. Process measures often become substitutes for outcome measures because they provide comforting evidence of activity, they demonstrate that *something* is happening.

Typically, both grantmakers and grantees contribute to process creep. It happens in the early stages of program implementation, when everyone involved suddenly becomes afraid that his or her hopes for the project may not be realized. It also happens when funders encounter hostility to outcome accountability (and outcome evaluation) from communities and program people who fear that outcome measurement will not do justice to their underfunded intervention.[13]

In responding to these fears, funders often find it easier to remove or move the goal posts than to strengthen the players. The typical forget-about-the-goal-posts conversation takes place a few months into the implementation phase of a program. The funder says to the grantee something along the following lines: So we gave you the grant in the hope that you would reduce teenage pregnancy and youth violence in this community, and you now say that was really an unrealistic expectation? You may be right. But we do need some hard evidence that our grant is making some sort of difference, so let's see if we can get an evaluator to design an attitude survey that will determine whether you have increased the number of teenagers who think it's a

bad idea to carry a gun and to initiate sex when they're younger than fifteen. Or the evaluators could document how many youngsters come to your meetings and classes. Alternatively, maybe we or you could hire an enthnographer to chronicle what's going on in your program ....

Some of these are useful things to do. It is especially useful to obtain rich descriptions of complex, nuanced interventions. But descriptions of process are most useful when they become part of a systematic inquiry into what the program is accomplishing and why. Descriptions of a process are not a substitute for either outcome-based accountability or outcome based evaluation.

Section 9: Conclusion

In concluding, I would address those who still harbor grave doubts and a visceral unease about the whole idea of results accountability. Committed practitioners have every reason to ask why should we have to prove the value of our work? They point out that those who would dismantle the safety net and the whole infrastructure of public and nonprofit services and institutions are not arguing efficacy—they are arguing principle. These practitioners, along with parents, community leaders and other advocates wish to stand their ground on principle, and say that feeding young children and providing them with a safe and happy place to play is enough justification, that comforting a frightened adolescent needs no further rationale, that every expectant mother is entitled to the highest quality prenatal care—regardless of whether there is a payoff in higher rates of school readiness, employability, or healthy births. Other countries, after all, do not make public support for basic services for children and families contingent on proof of their merit. In France and Germany and Britain and Japan, publicly supported child care and maternal and child health care, paid family leaves, and universal child protective services are taken for granted and require no evidence of effectiveness.

American human service leaders see themselves as part of a tradition of service to the vulnerable whose value is ultimately independent of its effects. They cite Mother Theresa's explanation of her perseverance in the face of the enormity of world poverty: "God has called on me not to be successful, but to be faithful" (Kagan, 1993). They cite Ghandi's teaching that "It is the action, not the fruit of the action, that is important."

My own belief is that the moral underpinnings for social action, especially by government, are not powerful enough today, in the cynical closing years of the twentieth century, to sustain what needs to be done on the scale at which it needs to be done. In this time of pervasive doubt, the magnitude of public investment that is required will be forthcoming only if there is evidence that investments are achieving their purpose and contributing to long-term goals that are widely shared. And the chances of developing and sustaining the responsive bureaucracies that can support effective programs will also increase to the extent that accountability for results can replace accountability for observing rigid and narrow procedural rules.

*Based on a chapter from a forthcoming book, tentatively titled *Disturbing the universe: Strong families, supportive communities, responsive bureaucracies, and how to get there from here.*

1. I use the words "results" and "outcomes" interchangeably. I use the word "goals" to refer to results that are desirable but cannot be readily measured or agreed upon.

2. While the bottom line of profit, and market performance and survival is clearly established in private business, there is no similar agreement on success in the public sector.

3. "You're seeing it everywhere," says James R. Fountain Jr., asst research director at the Governmental Accounting Standards Board, "a growing frustration among taxpayers that they don't know what they're getting for their money" (Osborne & Gaebler, 1993, p. 140).

4. President Bill Clinton's remarks at the signing of the Government Performance and Results Act: "It may seem amazing to say, but like many big organizations, ours is primarily dominated by considerations of input—how much money do we spend on a program, how many people do you have on the staff, what kind of regulations and rules are going to govern it; and much less by output—does this work, is it changing people's lives for the better?"(*From red tape to results*, 1993, p. 73)

5. Dr. Rivlin fully recognized the difficulty of coming up with good measures of performance: she called for "a sustained effort to develop performance measures suitable for judging and rewarding effectiveness ... all the strategies (here discussed) for finding better methods of delivering social services depend for their success on improving performance measures (Rivlin, 1971, p. 144).

6. "Ethical core" is Sid Gardner's phrase (Gardner, 1995). Gardner, along with David Hornbeck, has been the most persistent and effective advocate of the shift to an results orientation.

7. Robert Slavin reports that students who are not reading at the end of first grade are at great risk; they don't catch up later (Slavin, 1995).

8. School readiness can be thought of as "a fixed standard of development sufficient to enable children to fulfill school requirements and to absorb the curriculum content" (Kagan cited in Phillips & Love, 1994). The National Head Start Association says that Kindergarten teachers expect that entering students will be able to work both independently and as members of small and large groups, to attend to and finish a task, listen to a story in a group, follow two or three oral directions, take turns and share, care for their belongings, follow simple rules, respect the property of others, and work within the time and space constraints of a school program (National Head Start Association, 1995). Businessman and philanthropist Irving Harris tells the story of Doris Williams, who taught kindergarten in the inner city of Chicago, who told him she could always handle one child who wasn't ready for school. "But when I had two or three who were not ready the extended attention they demanded meant that the rest of the class was denied the time they had a right to expect from me" (Harris, 1993).

9. The NGA's community capacity indicators (benchmarks) for school readiness include rates of: children in preschool and child care

programs; children in preschool and child care programs meeting prescribed standards; eligible children in Head Start and public preschool programs; communities with family support and education services; school-age parents receiving comprehensive services; children who experience consecutive or multiple out-of-home placements; pregnant women receiving prenatal care during first trimester; children covered by Early Periodic Screening, Diagnosis, and Treatment (EPSDT) or private health insurance; eligible participants in the Women, Infant, Children program (WIC).

10. In an earlier paper, Love wrote, "Our system (in which assessments are completed on a representative sample of children) avoids labeling children by focusing on aggregate measures for the community."

11. In the spirit of Sarason's conclusion that, "Problems are constants, answers are provisional" (Sarason, 1990, p. xii), I would say that we could transform an understanding of our problems into an agreement about goals, and let those be our constants, as we evolve and experiment with our provisional programmatic answers.

12. In this process, it is important to start, as Alice Rivlin (1971) advises, with "indicators that measure movement in the appropriate direction." These include the following: measures of physical health (such as low rates of low birthweight babies, high rates of two-year olds fully immunized, no untreated vision or hearing problems at school entry, low rates of sexually transmitted diseases); measures of school achievement; measures of perils avoided in adolescence (such as too early childbearing, arrests for violent crime, suicide, homicide, substance abuse); measures of productivity and economic well-being (such as rates of productive employment, and rates of families with incomes over the poverty line) (Rivlin, 1971, p. 47).

13. The authors of *Outcome Funding* are troubled by the fact that a public college they describe was saved from closing by a campaign to preserve the jobs that the university provided in a poor county; they suggest that if that is one of the desired results, 50% of the school's budget should be funded by the state's economic development agency. Similarly, food stamps seems to be have been saved from the Gingrich assault in 1995 by agricultural interests, not by concern for the nutrition needs of the poor.

14. Brandon distinguishes between measures of wellbeing (which he calls positive measures) and measures of progress in reducing problems (which he calls negative measures) to argue in favor of using the former as the best way to approach results accountability. But he recognizes that "the popularity of negative measures—measures of poverty, dysfunction, and illness—reflects the ease of consensus on recognizing that some things are clearly inadequate, without the difficulty of consensus on defining what is adequate" (Brandon, 1992, p. 24).

15. I was at a meeting recently where Marie McCormick cited a study currently using IQ to find out how well a family support program was working.

16. In the vanguard of the work on neighborhood level indicators is the Foundation for Child Development, and researcher Claudia Coulton at Case Western Reserve University.

17. As one dramatic illustration, the coordinator of the external review team commissioned by the Pew Charitable Trusts to review its ambitious proposed Children's Initiative, suggested that the problem of obtaining timely results information contributed to the decision to cancel the Initiative. Reflecting back on lessons learned, he wrote that, "To build the public and political will to continue, projects must be able to demonstrate results in a credible way. State officials developing The Children's Initiative recognized that data on enhanced results were critical to expansion, but such data were unlikely to be available within the important early years of the project" (Krauskopf, 1994).

18. The evaluation steering committee of the Aspen Roundtable on Comprehensive Community Initiatives has been discussing the usefulness of a "Michelin Guide" to interim indicators, that would assess the degree of confidence with which the hypothesized connection between interim indicators and long-term results measures could be linked, all along the causal chain. The idea would be to distinguish among the connections that seem to be fairly well established, those where the evidence is weaker and the hypothesized connections urgently need to be tested, and those where even promising hypotheses are lacking.

19. Procedural protections will have to be maintained and monitored wherever there is no other way to restrict the arbitrary exercise of front-line discretion by powerful institutions against the interests of powerless clients. Because the present capacity to use outcome measures to judge program effectiveness is still primitive, and because it takes so long for results to improve in response to even the most effective interventions, existing process measures will continue to play a role in holding agencies, communities, and systems accountable during the period of transition. Increasingly, however, as new measures that are more closely and reliably related to results become available to measure initial progress toward ultimate goals, it will be important to continually re-examine the balance between the use of process and outcome measures, so that communities and agencies can make sure they utilize results-based accountability as much as the state of the art allows.

20. People who are responsible for programs, be they teachers, social workers, early childhood people, youth workers, or neighborhood residents and other program participants, often view evaluation research as an "unfriendly act," observed Peter Bell, when he was president of the Edna McConnell Clark Foundation (Bell, 1993).

References

Action Team on School Readiness. (1992). *Every child ready for school.* Washington, DC: National Governors' Association.

Barzelay, M. (1992). *Breaking through bureaucracy: A new vision for managing government.* Berkeley, CA: University of California.

Bell, P. (1993). Edna McConnell Clark Foundation report of the president.

Boyer, E. L. (1991). *Ready to learn: A mandate for the nation.* Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Brandon, R. N. (1992). *Social accountability: A conceptual framework for measuring child/family well-being.* Washington, DC: Center for the Study of Social Policy.

Bredekamp, S. (1995, May 8). Memorandum to S. L. Kagan, M. Levine, & V. Washington.

Campbell, D. T. (1987). Problems for the experimenting society in the interface between evaluation and service providers. In Kagan, S. L., Powell, D. R., Weissbourd, B., & Zigler, E. (Eds.), *America's family support programs: Perspectives and prospects.* New Haven, CT: Yale University Press.

Carnegie Task Force on Meeting the Needs of Young Children. (1994). *Starting points: Meeting the needs of young children.* New York: Carnegie Corporation of New York.

DiIulio, J. J. (1994). *Deregulating the public service: Can government be improved?* Washington, DC: Brookings Institution.

From red tape to results: Creating a government that works better and costs less. Washington, DC: National Performance Review. September 7, 1993.

Gardner, S. (1995, April). *Kissing babies and kissing off families.* Speech given at Concord, MA.

Harris, I. B. (1993, Spring). "Education—Does it make a difference when you start?" *Aspen Institute Quarterly, 5*(2), 30-52.

Kagan, S. L. (1993, May). *Going-to-scale with a comprehensive services strategy.* Presentation at Wingspread conference.

Kagan, S. L. (1995, May). *By the bucket: Achieving results for young children.* Washington, DC: National Governor's Association.

Krauskopf, J. A. (1994, October). *Overcoming obstacles to implementing reform of family and children's services.* Paper prepared for the annual meeting of the Association for Public Policy Analysis and Management, Chicago, IL.

Love, J. M. (1995, June). *Can we measure the results? Or, as society shifts towards results-based programs and services for young children, can the scientific community provide reliable, valid, and useful child results?* Paper prepared for discussion at the Issues Forum on Child-Based Results, sponsored by the Carnegie Corporation of New York.

Manno, B. (1994, June). *Outcome-based education: Miracle or plague?* Indianapolis, IN: Hudson Institute, Hudson Briefing Paper No. 165.

Moore, K. (1994, November). *Criteria for indicators of child well being.* Background paper for the Conference on Indicators of Children's Well-Being, Washington, DC.

National Head Start Association. (1995, May). *National Head Start Association statement on child-based results.* (1995, May). Washington, DC: Author.

National Research Council, Institute of Medicine. (1991). *Improving instruction and assessment in early childhood education.* Washington, DC: National Academy Press.
National Task Force on School Readiness. (1991, December). *Caring communities: Supporting young children and families.* Alexandria, VA: National Association of State Boards of Education.

Osborne, D. & Gaebler, T. (1993). *Reinventing government.* New York, NY: Penguin.

Phillips, D. & Love, J. (1994). *Indicators for school readiness, schooling and child care in early to middle childhood.* Paper prepared for conference on Indicators of Children's Well-Being, Washington, DC.

Rivlin, A. M. (1971). *Systematic thinking for social action.* Washington, DC: Brookings Institution.

Sarason, S. B. (1990). *The predictable failure of educational reform.* San Francisco, CA: Jossey-Bass.

Slavin, R. (1995, May). *Three days in May: How to assess first grade reading, and why we must do so.* Memorandum to May 1995 Carnegie Issues Forum on Child-Based Results.

White, S. (1995, June). Presentation at Carnegie Corporation Issues Forum on Child-Based Results, New York, NY.

Young, N. K., Gardner, S. L., Coley, S. M. (1993). *Getting to results in integrated service delivery models.*

Zill, N. (1995, May). *Responses to Issues Forum Meeting questions.* Carnegie Issues Forum on Child-Based Results, New York, NY.

Appendix E: *Can We Measure the Results?*

Or, As Society Shifts Toward Results-Based Programs And Services For Young Children, Can The Scientific Community Provide Reliable, Valid, And Useful Child Outcome Measures?

John M. Love
Mathematica Policy Research, Inc.

Paper presented at the
Issues Forum on Child-Based Results
New York City

Section 1: Introduction

Lisbeth Schorr (1994) has made the case for shifting to results-based accountability as we strive to improve the lives of children and families. I accept the desirability of this shift. But now what? Can we actually measure the results or results that programs are now clamoring to articulate? I have been asked to consider this question from the perspective of science—the science of child development and the science of measurement.

WHERE THERE'S A WILL THERE'S A WAY

At some level, the answer has to be "yes," because we continually measure a wide range of results in a large number of programs. In fact, there is a long—though some might say checkered—history of using existing instruments to create a body of knowledge on the effectiveness of programs for children. Head Start, for example, has long measured its effectiveness with a variety of child outcome measures (Kresh 1993; McCall 1993; McKey et al. 1985). Studies of other preschool programs, like the Perry Preschool, have influenced public policy in major ways with only minimal questions about the accuracy of the results measures (Barnett 1992; Schweinhart, Barnes, and Weikart 1993). Early intervention studies, like the Infant Health and Development Project (IHDP) (Brooks-Gunn et al. 1994) and the Abecedarian project (Campbell and Ramey 1994; Ramey and Campbell 1991), are currently claiming significant benefits based on existing child outcome instruments. Extant instruments form the basis for our understanding of the benefits of high-quality child care (Helburn et al.1995; Howes, Phillips, and Whitebook 1992). Furthermore, studies of family support programs also use existing measures to assess results for children (Barnett 1995; Larner 1992; St. Pierre, Layzer, and Barnes 1994; Yoshikawa 1994). Because we are measuring child results, there is at least tacit acceptance of the outcome measures by a sizable group of researchers, program operators, and policymakers. Without this acceptance, we could not have any confidence in the hundreds of studies we rely on for our understanding of programs, their accomplishments, and their impacts on children.

If, at some level, many researchers and program planners believe we can currently measure important results, we should still ask whether we are doing so as well as we should. Could we use better measures? Are we measuring the right results? Or the most important ones? Do we fail to learn about particular kinds of program effects because we lack the instruments? With sufficient resources—that is, time and money—wisely applied, there is no doubt in my mind that the scientific community can answer the feasibility question affirmatively. But, for this paper, I will assume there is no time and no additional money. The sponsors of the forum are interested in what is feasible now, not what is eventually possible!

Schorr (1994) and her colleagues did consider measures of programs' results for children that are suitable for immediate use. Her "start-up list" of possible measures includes many of the usual suspects—lower rates of low birthweight babies, more complete immunizations, lower school dropout rates, and fewer children abused or neglected, to name a few. These are critical results, particularly from a cost-conscious public policy perspective. Conspicuously absent, however, are measures of the developmental results that many early childhood program people—teachers, caregivers, parents, and child development experts—care about: communication skills, thinking ability, self-esteem, school-related knowledge, behavior problems, curiosity, and so forth—the multiple dimensions of children's early learning and development that have been so admirably and comprehensively articulated by the technical working group for the first national education goal (Kagan, Moore, and Bredekamp 1995). Although aspects of these developmental domains are commonly measured, as they have been in the studies cited earlier, we could long debate their scientific suitability.

In this paper, I discuss the scientific feasibility of measuring a wide spectrum of child results. All are encompassed by the notion of children's well-being. All are articulated among the goals of many programs that are striving to improve results for children and families. I illustrate measurement of various facets of well-being selected from the typical domains: health and physical development, intellectual or cognitive functioning, language and communication, and social and emotional development.

It may well be that measurement is more feasible in some domains than others. I also consider the possibility that feasibility is a function of the age of the child. I raise the possibility that measurement feasibility depends on other characteristics of the child, such as his or her native language and culture, and special needs or disabilities. Feasibility may vary with the type of measurement as well. Finally, I suggest that the science of psychological measurement is not the only factor influencing the feasibility of adopting an results-based orientation for children's programs. In fact, in spite of the theme of this forum, the scientific feasibility of measuring child results may not be the most important source of apprehension. Before getting to this topic, however, I discuss some of the more salient past and present child-outcome measurement efforts.

Section 2: Past Assessment Efforts: Heritage or Hamstring?

Within the early childhood community, the infamous "Westinghouse" study stands as a hallmark of failure in the science of evaluation—failed design, failed measurement, and failed policy. Not only did this early evaluation concentrate on measuring the intellectual development of Head Start children with an inappropriate test of "IQ" (chosen because it was "the best measure available"), it included noncomparable comparison groups under noncomparable program conditions. Although design flaws, unrealistic expectations, and problems with the media (including premature release of findings to *The New York Times*) compounded a bad choice of measures, it is often the outcome measure that takes the heat of later condemnation.

If the so-called Westinghouse evaluation set the field back a decade, the long-term benefits shown for graduates of Ypsilanti's Perry Preschool immeasurably advanced the cause of early childhood programs. I don't think there is a single study that has so inclined politicians to vote for additional funds for Head Start. In this case, a randomly assigned control group, long-term follow up, and results that really matter (getting jobs, staying out of jail, avoiding pregnancy) far outweighed earlier reports of fading IQ gains (Barnett 1992). In fact, the Perry Preschool study, along with the other longitudinal early childhood education studies begun in the 1960s (Lazar, Darlington, Murray, Royce, and Sniper 1982), were instrumental in awakening interest in a wide range of educational, social, and economic results that extend well beyond typical measures of children's development.

Another program from this era was the Brookline Early Education Project (BEEP), distinguished by being an integral part of the public school system and providing services to children and families from birth until entry into kindergarten. Outcome measures included standard developmental measures (such as the Bayley Scales of Infant Development and the Stanford-Binet); measures of language development, health and physical development, and school achievement; and a detailed classroom-observational measure of children's mastery skills, social skills, and use of time (Hauser-Cram, Pierson, Walker, and Tivnan 1991). The broader range of results measured for BEEP children can be justifiably envied by today's early childhood programs.

This highly selective sampling of early childhood program evaluations from the '60s and '70s shows that the field did not lack for outcome measures. Yet, study after study has begun by declaring that adequate measures do not exist. There are sound scientific and political reasons for not wanting to perpetuate the "psychometrically" strong IQ tests, and the highly program relevant and sensitive, but labor-intensive, observational measures like those used in the BEEP evaluation cannot easily be used on a large scale.

In the context of these exciting studies, about twenty years ago, I began planning the evaluation of a Head Start demonstration program that had ambitious goals for affecting 4- to 8-year-olds' "social competence." The evaluation team identified four dimensions of social competence in an effort to do justice to Ed Zigler's broad conception of social competence as "an individual's everyday effectiveness in dealing with his environment, . . . his ability to master appropriate formal concepts, to perform well in school, to stay out of trouble with the law, and to relate well to adults and other children" (quoted by Anderson and Messick 1974, p. 283). The early 1970's precursor to the Administration on Children, Youth and Families (the Office of Child Development) simplified the concept to a concern with the child's "everyday effectiveness in dealing with his environment and responsibilities in school and life." The domains we identified were social-emotional development, psychomotor development, language development, cognitive skills, and health and nutrition—not unlike the current dimensions defined by the Goal 1 Technical Planning Group of the National Education Goals Panel (Kagan, Moore, and Bredekamp 1995).

In 1975, we concluded that "no measure that is already fully developed has been found that meets all the specific selection criteria . . ." (Love, Wacker, and Meece 1975, p. 3). It is instructive to look at the instrument selection criteria used in that study, because they are similar to those followed in practically every early childhood program evaluation I know of, and are still relevant to the issue facing us today. Fifteen criteria in three areas were used to evaluate potential instruments:

Practical Considerations:

- Available for use by fall 1975 ("immediately")
- Appropriate for use by trained paraprofessionals
- Test format appropriate for ages of children in program

BEST COPY AVAILAB

- Scoring procedures appropriate for data processing
- Reasonable testing time for young children

Psychometric Qualities:

- Adequate construct and/or predictive validity
- Adequate test stability and internal consistency
- Culture and/or SES fair
- Representativeness of standardization sample
- Low correlation with index of general information

Relevance to the Program:

- Spans appropriate age range
- Spanish-language adaptation available
- Relevant to program's cognitive and language goals
- Likely to demonstrate program effects
- Used in previous national evaluations or large-scale studies

Each of these criteria represents a factor influencing our answer to the question of scientific feasibility of measuring policy-relevant child results. In this instance, the primary shortcomings of the instruments were their failure to span the total 4- to 8-year age range of the program population, lack of relevance to program goals, and failure of the test standardization samples to represent fully the geographic or SES features of the population participating in the program.

During this same period, the Office of Education (OE), the middle element in what was then the U.S. Department of Health, Education, and Welfare, was conducting a massive national evaluation of an early elementary (kindergarten through third grade) planned-variation curriculum study called "Follow Through." In addition to conducting its multimillion-dollar national evaluation, OE allowed the curriculum sponsors to conduct their own evaluation activities. A consortium of the institutions sponsoring curricula that today we would call "developmentally appropriate" joined forces to develop measures that would be more sensitive to developmental results than the standardized achievement tests used in the national evaluation.

The High/Scope Follow Through model, for example, developed a procedure for assessing written language that demonstrated that fluency and complexity in the writing of Follow Though students increased as a direct function of the extent to which teachers implemented the High/Scope curriculum (Bond, Smith, and Kittel 1976). The Bank Street College Follow Through model developed a complex observation instrument that showed Follow Through children engaging in more self-initiated communication, expression of thoughts, and peer communication than comparison children (Bowman and Mayer 1976). These and other sponsor evaluation efforts were well-intentioned and in fact did produce useful findings to counterbalance the national evaluation results (Hodges et al. 1980). But the enormous task was too much for the sponsors' paltry evaluation resources and time, and they could not effectively counter the negative messages of the national evaluation.

In the late 1970s, OCD decided that the sorry state of measurement feasibility was a major obstacle to obtaining good evaluations of the Head Start program. The agency therefore launched a coordinated effort to develop new instruments. The strategy included contracts to Educational Testing Service and the RAND Corporation to define and map the landscape of what should be measured (Anderson and Messick 1974; Raizen and Bobrow 1974), and two major instrument development projects. In the first, OCD contracted with ETS to create a battery of measures spanning 16 domains—in the areas of cognitive, language, and perceptual-motor characteristics of preschool and kindergarten children. Called the CIRCUS battery because of the pictorial theme using clowns, balloons, and animals, the battery offered a special feature, EL CIRCO. This Spanish edition was distinguished by the fact that it was developed in tandem with the English-language version rather than being a translation of a test initially developed only for English-speaking children (Anderson et al. 1974). Today, no one hears about CIRCUS or EL CIRCO.

The second project was the Head Start Measures Project, begun in 1977. In 1980, the project team conducted an extensive review of measures of children's development. We reviewed more than 200 measures, reaching the following conclusion:

> Very few measures show content validity as defined by the developmental characteristics of children identified through the input workshops. Information on construct validity is virtually nonexistent. Moreover, very few measures possess sufficient reliability to warrant confidence in evaluation findings based on their use. (Mediax Associates 1980, p. 34)

Members of the national panel convened for the Head Start measures project had equally discouraging observations. The comments of panelists representing three perspectives illustrate the measurement problem that we, and the field, faced:

> I have reviewed all of the major studies of Head Start and related programs and all of the instruments used in this research to assess children's social development. I cannot recommend any of these instruments for adoption in this project since in no case did I find satisfactory evidence of the instrument's validity. (Carew 1978, p. 7)

> Indeed, most of the scales of motor development available contain so few items per age level . . . that they are unlikely to be discriminating except under circumstances of marked differences between the groups being assessed. (Eichorn 1978, pp. 5-6)

> It is clear that evaluators of Head Start have not taken into account, in their selection of measures, the com-

plex issues underlying the identification of goals and objectives of the program. (Laosa 1978, p. 19)

The Head Start Measures Project is the only sustained effort I know of that was single-mindedly devoted to developing better outcome measures for the full spectrum of children's well-being. Researchers across four different institutions tackled instrument development in four domains: (1) health and physical; (2) cognitive; (3) social-emotional; and (4) "applied strategies." (This fourth domain foreshadowed the "approaches toward learning" domain defined by Kagan et al. [1995] for the first national education goal.) After only two years of development work, the government canceled funding for all but the cognitive domain. Unfortunately, the project's dismal outcome, documented by Raver and Zigler (1991), does not make us eager to try again.

Today, we are often hamstrung by failed measurement development attempts and the negative exemplars of major evaluations of early childhood programs. We are told that this history shows it cannot be done. These experiences are at least partly responsible for a climate of measurement avoidance throughout the early childhood community. It is now "common knowledge" that the concept of test validity for young children is an oxymoron. Teachers, administrators, or program evaluators who recommend individual, standardized, controlled assessments of young children are misguided at best and evil at worst. I am not so naive as to believe that valid concerns do not exist, or to recognize that the practice of testing has involved enormous abuses, but it is both wrongheaded and shortsighted to reject all testing out of hand. Even within this climate, however, positive examples exist. I turn now to some of these to consider what lessons they carry.

Section 3: Recent and Current Assessment Efforts: Enlightened Endeavors or Imprudent Illusion?

In three recent and current activities, researchers have identified child results that are relevant for particular purposes and have proposed useful measurement approaches. Here, we see that different types of measurement procedures are identified; however, constraints on the definition or construction of the measures in each case cloud our ability to focus on the nature of the problem for results-based programs.

CHILD RESULTS IN THE CONTEXT OF COMMUNITY-BASED FAMILY PROGRAMS

In 1993, Mathematica Policy Research undertook the challenge of designing an evaluation of the Pew Charitable Trusts' Children's Initiative that would be as comprehensive and innovative as the Initiative itself. Although we knew the task would not be easy, we were convinced it was feasible. I choose The Children's Initiative evaluation as an example because it illustrates the child results we were convinced could be measured within the context of a community-wide program that also had broad health and family functioning goals. When The Children's Initiative ended, Thornton, Love, and Meckstroth (1994) extended Mathematica's design work to propose a system of measures that would be appropriate for other community programs with outcome goals similar to those of The Children's Initiative.

We developed plans for measures related to results in five broad areas: (1) child and family health; (2) family functioning; (3) child development; (4) school performance; and (5) youth maturation and social integration. Two of the topics under child and family health (incidence of preventable diseases and disabilities, and overall health of children), as well as the areas of child development and school performance, are pertinent to the child-based results theme of this forum. For each outcome that might relate to a goal of a community program in each of these areas, we listed the expected results (both intermediate and long term), the recommended measure for each outcome, and the measurement procedure or data source. We also provided a summary of evidence, when available, on how sensitive the measure is to community interventions, the strengths and liabilities of the measure, and a brief statement as to its policy relevance (see Tables 1 through 4).

As with any set of recommended measures, we developed this list with some restrictions in mind. First, we tried to find measures that are sensitive to changes in the well-being of children brought about by community-wide programs. Second, we wanted measures based on data that local communities would be capable of collecting. Third, we gave priority to measures for which there would be comparative data in national surveys or other studies that could provide a frame of reference for interpreting local data. Finally, our selection was influenced by the potential for making policy relevant statements about changes in children's status on the measure.

The measures in Tables 1-4 rely on two major data collection strategies: interviews or surveys and administrative records. We did not include controlled, individualized assessments of children—not because of concerns about their scientific feasibility, but because of practical feasibility, primarily the costs of data collection. A major reason for constraining our selection of measures was our desire to make available a system that communities could maintain after the evaluation ended. Communities seldom have the resources for ongoing assessments that are possible with a specially funded formal evaluation. We thought we were proposing a scientifically feasible system of measures, appropriate for the goals (desired results) of The Children's

Initiative and suitable for implementation in the context of community-wide programs, but six months after we submitted our preliminary evaluation design, the Trusts decided not to proceed with implementing the Initiative.

I have often wondered how large a role the adequacy of the outcome measures played in the Trusts' decision. Clearly, it was a complex decision in which many factors were weighed. As the coordinator of the external review team commissioned by the Trusts to examine the Initiative objectively, Krauskopf (1994) has reflected on the lessons learned. One of the four areas of concern he articulated was "evaluation, outcome measurement, and accountability." Referring to the outcome focus of the Initiative, Krauskopf noted that "there are not well-agreed upon indicators and measures for three of the Initiative's four outcome areas—child development, family functioning, and school readiness." Krauskopf's central concern in the measurement area, however, makes it clear that the adequacy of the outcome measures is a concern, but that this concern is embedded in a larger matrix of issues that includes evaluation design and the policymaking process:

> Because large-scale systems projects must be prepared to justify themselves as they proceed, the *absence of solid outcome measurement data* and operational management systems is particularly harmful to generating ongoing support. To build the public and political will to continue, *projects must be able to demonstrate results in a credible way.* State officials developing The Children's Initiative recognized that *data on enhanced results* were critical to expansion, but such data were unlikely to be available *within the important early years of the project.* (Emphases added)

In part, we may have failed to demonstrate the scientific feasibility of producing the valid measures outlined in Tables 1-4; certainly, none of the measures is perfect. It must be recognized, however, that the scientific basis for outcome measurement, even had it been extremely strong, would not have been sufficient. Two other factors are also important. First, programs must be able to translate the results of measurement into policy-relevant conclusions about the linkage between program activities and measured results: Did this specific community intervention actually produce the changes that evaluators observed in the performance of children on the measures? Second is the issue of timing. There is almost always a tension between the research and evaluation schedule, which must wait for the program processes to unfold and produce their impacts, and the political agenda, which demands immediate "hard" evidence that the investment is paying off. In this context, as we have seen in the case of Head Start, it is often tempting to blame the measures.[5]

CHILD RESULTS IN THE CONTEXT OF SCHOOL READINESS ASSESSMENT

My second example takes us to a particularly sensitive arena, that of measuring the extent to which the developmental progress of 5-year-old children provides them the wherewithal to succeed in school. The issue of school readiness assessment is rife with debate, the pros and cons of which I will not go into here. Let it suffice to say that the attitude of measurement avoidance is particularly acute when the results of the measurements have the potential for labeling children and leading to critical decisions about their educational trajectories (grade placements, tracking, and so forth). With support from the Pew Charitable Trusts, Larry Aber, Jeanne Brooks-Gunn, and I analyzed the dimensions of children's early learning, development, and abilities as defined by the Goal 1 Technical Planning Group (Kagan et al. 1995) and searched the literature for measures that would be appropriate for communities to use to provide data on important elements of each dimension. The results of this process are summarized in Table 5.[6]

Each of the measures listed in the right-hand column of Table 5 has been used in previous surveys or research and is supported by evidence of its reliability and validity. Also important for the question of feasibility, we judged each measure to be appropriate for diverse groups of children representing the entire spectrum of socioeconomic status, geographic regions, racial-ethnic groups, language groups, and disability status found in this country. Each of the measures was selected because it assesses one or more of the constructs embodied in one of the dimensions of children's development and learning, is appropriate for 5-year-olds, and is available for use with relatively little further adaptation. Two other considerations further constrained our search for appropriate measures. First, we wanted the set of measures, taken as a whole, to be practically feasible. That is, the assessment process should not be so time-consuming or require such extensive training or materials that school or community personnel would have difficulty administering the measures on a relatively large scale. Second, we looked for measures on which we had access to national data for drawing comparisons.

Any application of child results measurement will have a similar set of constraints. We must keep these in mind as we contemplate the scientific feasibility of assessing child results. In other words, the characteristics of the measures are not the only consideration.

Four separate measurement procedures are required to assess all these indicators: (1) a self-administered parent questionnaire; (2) teacher-conducted assessments of children's development; (3) ratings by teachers; and (4) school health records. Feasibility is enhanced by the use of multiple sources of data. For example, both parents and teachers

complete ratings on aspects of children's social and emotional development. Teachers obtain data on children's motor development, approaches toward learning, language usage, and cognition and general knowledge using a standardized screening instrument.

This system of measures is yet to be tried, but it appears to have face validity, if the expressions of community interest that we have received are any indication. We know that the measures exist and are scientifically strong. We don't know the full extent to which (1) their administration will be practical and affordable; or (2) the data they generate will speak to the needs of communities that are concerned about these dimensions of development as possible results of their systems of health care, child care, parent support, and early education.

CHILD RESULTS IN THE CONTEXT OF A SYSTEM OF INDICATORS

Last year, the organizers of a conference on "indicators of children's well-being" commissioned 24 papers, in which prominent researchers described what child and adolescent well-being indicators are feasible to measure. Recommendations were subject to a constraint unlike those seen in the previous examples: the measures must be obtainable from national surveys. Even with this severe restriction, conference participants considered a large number of measures of children's well-being to be both feasible and desirable to collect on a national basis. The measures cover the domains of child health; education; economic security; population, family, and neighborhood; and social development and problem behaviors (Brown 1994). The following partial listing from that conference illustrates what is feasible within the context of national survey data collections for children from birth to 5 years:

Child Health
- Healthy birth index
- Percentage of infants born with congenital anomalies
- Child abuse/neglect rate
- Percentage of children ever experiencing a delay in growth or development
- Percentage of children limited by chronic health conditions
- Percentage of children who regularly use seat belts

Education
- Percentage of 3- to 5-year-olds enrolled in preschool
- Percentage of children ages 3 to 5 who are read to every day by a parent or household member
- Percentage of children over 3 years who ever had learning disabilities

Economic Security
- Percentage of children in poverty
- Percentage of children in families receiving food stamps in past year
- Percentage of children in households where both or only parents are working
- Percentage of children living in inadequate housing

Population, Family, and Neighborhood
- Percentage of children who have moved within the past year
- Percentage of children living in institutions or group quarters
- Percentage of children living in severely distressed neighborhoods

Social Development and Problem Behaviors
- Percentage of children with high rates of behavior problems

Because these measures must be collectable through a national survey, they may be less useful for measuring results in certain types of programs. On the other hand, they may be particularly useful for evaluating community-wide programs like The Children's Initiative and many of the family support programs around the country.

Section 4: It all Depends: Tentative Lessons About Influences on Measurement Feasibility

Three different strategies have illustrated what some researchers believe to be the scientific feasibility of measuring child results. To make this discussion very concrete, I now describe four illustrative measures. These examples have practical feasibility, which I took as a threshold requirement for considering any measure useful to this forum. In other words, either they have been used in large-scale studies, where the cost and burden of data collection are important considerations, or I can imagine them being used in such studies. I have selected these four to illustrate that (1) outcome measurements *are* feasible; (2) a range of features of children's development and well-being *can* be measured; but that (3) a number of factors influence measurement feasibility. They do not constitute a random sample of measures, but neither do I think they are entirely atypical. I realize that to fully answer the question of this forum, this analysis should be done with a much larger sample of measures. If, however, these four suggest a positive answer, this analysis should provide sufficient grist for our debate.

Table 6 summarizes critical features of the four instruments: (1) Behavior Problems Index (BPI) (Zill 1990); (2) Early Screening Inventory (ESI) (Meisels et al. 1988); (3)

BEST COPY AVAILA

MacArthur Communicative Development Inventories (CDI) (Fenson et al. 1993); and (4) Social Skills Rating System (SSRS) (Gresham and Elliott 1990). The results measured by these four instruments include aspects of many of the important dimensions of children's early development and learning: motor, social-emotional, language, and cognition and general knowledge. The instruments span different age ranges, from infancy through the elementary grades. They also represent different measurement approaches: parent reports and ratings (CDI and BPI); teacher ratings (SSRS); and direct individual assessments of children in controlled settings (ESI). All but the CDI have been relatively widely used. The CDI is still undergoing developmental research (Fenson 1993), whereas many large-scale surveys have incorporated the BPI (Love 1994; and Zill and Schoenborn 1990). One of the measures (CDI) was explicitly designed to measure results for infants and toddlers; the BPI is probably suitable down to age 2, even though it was initially developed for preschoolers and older children (Brooks-Gunn and Ross 1991). The ESI is designed primarily for the preschool-to-first-grade years, and the SSRS teacher version is designed for elementary school teachers to complete (preschool and secondary versions are also available).

I've indicated that it is not completely fair to draw broad generalizations from this small subset of measures, so I will overgeneralize! Using these four instruments, as well as knowledge about many other measures that is not documented here, I now suggest some of the factors that may determine the scientific feasibility of adopting an results orientation for early childhood programs.

AGE AND OTHER CHARACTERISTICS OF THE CHILD

The older the child, the stronger the measures. But not always! As a general rule, when children become older, they are easier to talk to, they better understand what we ask of them, they respond more consistently to instructions. If our measures depend on communicating with the child and getting a response in return, then this general rule applies. Standard aptitude tests are often thought to be more valid with older children and adolescents than with babies. When a good multidimensional developmental assessment is created (such as the Bayley Scales of Infant Development), it is rejected for all but the most well-funded evaluations. Notice, however, that such in-depth measures often become the benchmarks against which the validity of newer, more efficient assessments (like the CDI) are judged.

The very widely used Peabody Picture Vocabulary Test (PPVT-R) (Dunn and Dunn 1981) is a good example of an instrument that spans a very wide age range, is practical to administer, but is *inappropriate* for young children. This test relies on formalized interactions between adult and child, with the adult asking questions (for example, "Which is the ladder?"), and the child being required to respond in very restricted ways by pointing to or giving the letter designation for the correct one out of four pictures. There is ample room for bias to creep in—the adult's pronunciation, the familiarity of the vocabulary within the child's culture, the child's interpretation of the inelegant drawings, and so forth. In spite of these problems, the PPVT has been widely used in program evaluations, like the Infant Health and Development Program (IHDP), and large-scale surveys, like the National Longitudinal Survey of Youth (NLSY). It has a number of enviable psychometric characteristics, including good reliability and predictive validity, yet fails on our criteria of appropriateness for diverse cultural and racial groups (because of vocabulary selection and drawings) and relevance to program goals (since it measures only receptive vocabulary understanding—a very narrow slice of the goals that most programs have for child development). Language has always been a difficult outcome to measure, partly because of its sensitivity to context. So, perhaps the domain of development is more important than the child's age.

DOMAINS OF DEVELOPMENT

For the last 10 or 15 years, Larry Fenson and his colleagues have struggled with ways of measuring the *productive* language of infants and toddlers. What have they found? That parents are not only a convenient, but a highly reliable, source of information on the vocabulary and syntax of their own children. Parents (mostly mothers) can observe their child's language across multiple settings over time in ways that would be extremely time-consuming and expensive for outside observers to duplicate. Taken together, the infant and toddler scales of the CDI describe the course of language development, from nonverbal gestures in infancy, through early vocabulary usage, to the beginnings of grammatical speech. Interestingly, the growth curves fitted to the parent reports of various language functions closely parallel predictions from language theory (Dale, et al. 1989; and Fenson et al. 1993). So here we have the case of a strong measure for very young children in a domain that is fraught with measurement difficulties. Should we, perhaps, put more reliance on parent reports?

MEASUREMENT METHODS

The Behavior Problems Index provides another example of the successful use of parent reports of children's behavior. But the Social Skills Rating System has been more successful with its teacher-rating version than with the parent form, at least in terms of the internal consistency of the scales (Gresham and Elliott 1990). On the other hand, the Early Screening Inventory has almost single-handedly

erased our usual disdain for developmental screening instruments. The ESI demonstrates not only reliability and validity but also *sensitivity*, an especially important quality when teachers and other school personnel are concerned with correctly identifying children who should be referred for further diagnosis. It has not been widely used in program evaluations (and, in fact, was not developed for that purpose), but it is certainly a better candidate than many of the traditional screening instruments that have been so used.

In-person assessments, like the ESI, which rely on structured settings that carefully control the demand characteristics for children's responses, generally cost more to administer—if they go beyond picture-book multiple-choice formats. Contributing to the psychometric strength of the ESI are certainly the systematic procedures and training programs that ensure consistent administration across teachers and settings. Good parent and teacher rating forms, like the BPI, CDI, and SSRS, also provide systematic instructions for their administration, but that task is easier since direct contact with the child is not required. Thus, it seems that the measurement method is not a determining factor so much as the care with which procedures are established to ensure that the intentions of the instrument developer are carried out.

EVALUATION DESIGN ISSUES

Characteristics of the child, the domains being measured, and how we go about conducting the measurement process are all important considerations. It is not clear to me, however, that our experience suggests useful rules of thumb for selecting the best measure for a particular program's outcome goals. There is simply too much measure-to-measure variation, as seen in the perhaps extreme difference between the PPVT and the CDI. It is time to consider not only the measures themselves, but also how they are used.

Suppose we have one or two scientifically valid and reliable instruments that program stakeholders agree measure results they are trying to achieve. How, then, are the results of our scientifically valid outcome measures to be used? The desirability of moving toward a child-based results orientation is based on the assumption that the results of the measurement process mean something to the various program stakeholders—policymakers, early care and education systems people, families and communities, and program staff. Even with a perfectly valid measure of children's language usage or behavior problems, we cannot interpret the scores unless there is highly convincing evidence that the program had something to do with them. This is not the place for a thorough discussion of evaluation design issues, but we must recognize that even the best outcome measures are *useless* in the face of our inability to implement evaluation designs that permit unambiguous conclusions about program effects.' I submit that problems in evaluation design are much more likely than invalid outcome measures to be the "scientific" basis for impeding movement toward results-based programming for children.

Section 5: Closing Thoughts

Each year about 750,000 children enroll in Head Start. Responding to recent concerns about program quality, ACYF is designing a system of "performance measures" that will be results-oriented. The system will include measures of children's social competence, because the agency is convinced that quality improvements will be reflected in better results for children. Does the field have perfect measures to recommend to ACYF for use next year, or the year after? No. Should we tell ACYF, and indirectly the Congress, that it is impossible to judge the results of enhanced quality? I think not. We know enough now to specify the conditions under which it is scientifically feasible to measure child results of Head Start.

Three months from now, about four million children will enter public and private schools for the first time. Will they have a successful experience? Well, a sizable proportion will not actually enter regular kindergartens because their parents, school system, or private counselors will deem them "unready." Twelve months later, almost 200,000 children will spend a second year in kindergarten because their school doesn't believe they have made enough progress to be successful in first grade. An additional 160,000 or so will also experience delayed entry to first grade because their school system places them in some type of transitional class after the kindergarten year. Many of these decisions (50 percent or more of the cases) will be justified by the results of some type of assessment. We already know better ways to make these decisions, ways that will place children at far lower risk for school failure, later dropout, and so forth (Shepard and Smith 1989). Are any of these outcome measures perfect? No. But each year that we wait for the perfect measure, we diminish the chances for successful schooling for another 350,000 or more children.

Uncounted millions of children participate in family support programs each year at a cost of tens of millions of dollars. There is keen interest in knowing the extent to which these children are better off because of the support and services their families receive. Should we tell policymakers that it is premature to assess these benefits? That programs will just have to proceed without knowing how they affect children's well-being? Again, I say no. And again, we know enough to advise programs on the conditions under which effective results measurement can be conducted, on which dimensions of well-being, and with which measurement procedures.

If it is desirable to shift to results-based accountability in

BEST COPY AVAILABLE

programs for children, it is also scientifically feasible to do so. I am fully aware that this brief paper and my selective review of measures do not solidly bolster this conclusion. Yet, it seems impossible to deny the good measures we have available. In arriving at this conclusion, I want to be clear that an affirmative answer to the feasibility question is *not* a blanket endorsement of all measures for all children in all programs under all circumstances. The results that are measured, the procedures that are selected, and the conditions under which the assessments are conducted and interpreted all have implications for programs and families, and ultimately for the children whom we hope will benefit through our endeavors.

1. The planning phase of the Measures Project included a series of "input workshops," in which project staff met with representatives of Head Start administrators, teachers, and parents in various regions of the country. Conducted like large focus groups, these workshops were designed to ascertain what representatives of the Head Start community (the stakeholders of future program evaluations) believed to be important indicators of children's development.

2. Although Tables 1-4 have their foundation in preliminary recommendations Mathematica Policy Research developed for possible use in the evaluation of The Children's Initiative, we have made a number of modifications since The Children's Initiative ended. No endorsement of these particular results or measures by the Pew Charitable Trusts should be inferred.

I am grateful to Craig Thornton for helping me articulate the complexities of measurement issues in the context of evaluating community-wide initiatives.

3. For the purposes of this discussion of child results measures, we can ignore the first four dimensions in Table 5. They reflect the community conditions that the National Governors' Association identified as supporting children's development and learning and are important for a complete assessment within the framework of the first national education goal (see Love et al. 1994).

4. The conference was sponsored by the Institute for Research on Poverty at the University of Wisconsin; Child Trends, Inc.; the Office of the Assistant Secretary for Planning and Evaluation (U.S. Department of Health and Human Services); the National Institute of Child Health and Human Development; and the Annie E. Casey Foundation.

5. I will not go into the power of random assignment, the virtues of having a control-group counterfactual, or the various quasi-experimental design alternatives available. Hollister and Hill (1995) have recently presented a highly intelligent and readable discussion of these issues with particular reference to evaluating communitywide initiatives.

References

Anderson, Scarvia, and Samuel Messick. "Social Competency in Young Children." *Developmental Psychology*, vol. 10, 1974, pp. 282-293.

Anderson, S.B., G.A. Bogatz, T.W. Draper, A. Jungeblut, G. Sidwell, W.C. Ward, and A. Yates. *CIRCUS Manual and Technical Report* (preliminary version). Princeton, NJ: Educational Testing Service, 1974.

Barnett, W. Steven. "Benefits of Compensatory Preschool Education." Journal of Human Resources, vol. 27, 1992, pp. 102-120.

Barnett, W. Steven. "Research on Family Support Programs: Current Status and Future Directions." Paper presented at the biennial meeting of the Society for Research in Child Development, Indianapolis, March 1995.

Bond, James T., Allen G. Smith, and Janice M. Kittel. "Evaluation of Curriculum Implementation and Child Results." Annual Report, Cognitively Oriented Curriculum Project Follow Through. Ypsilanti, MI: High/Scope Educational Research Foundation, 1976.

Bowman, G., and Rochelle Mayer. "The BRACE System of Interaction Analysis: What, Why, and How? New York: Bank Street College of Education, 1976.

Bradley, R., and B. Caldwell. "The Relation of Infants' Home Environments to Achievement Test Performance in First Grade: A Follow-Up Study." *Child Development*, vol. 55, 1984, pp. 803-809.

Bradley, R., and B. Caldwell. *Home Observation for Measurement of the Environment: Validation Studies of the Preschool Version.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, 1978.

Brooks-Gunn, Jeanne, and Christine Ross. "Evaluating Child Results of the Expanded Child Care Options Demonstration: An Analysis Plan." Princeton, NJ: Mathematica Policy Research, Inc., May 1991.

Brooks-Gunn, J., C. McCarton, P. Casey, M. McCormick, et al. "Early Intervention in Low Birthweight, Premature Infants: Results Through Age 5 From the Infant Health and Development Program." *Journal of the American Medical Association*, vol. 272, no. 16, 1994, pp. 1257-1262.

Brown, Brett. "Indicators of Children's Well-Being: A Review of Current Indicators Based on Data From the Federal Statistical System." Paper prepared for the conference on Indicators of Children's Well-Being, Bethesda, MD, November 1994.

Campbell, Frances A., and Craig T. Ramey. "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families." *Child Development*, vol. 65, 1994, pp. 684-698.

Carew, Jean V. "Head Start Profiles of Program Effects on Children: Position Paper for Mediax ACYF Project." Westport, CT: Mediax Associates, 1978.

Dale, Phillip S., Elizabeth Bates, J. Steven Reznick, and C. Morisset. "The Validity of a Parent Report Instrument of Child Language at 20 Months." *Journal of Child Language*, vol. 16, 1989, pp. 239-249.

Dunn, L., and L. Dunn. *Peabody Picture Vocabulary Test, Revised Edition*. Circle Pines, MN: American Guidance Service, 1981.

Edozien, J., B. Switzer, and R. Bryan. "Medicaid Evaluation of the Special Supplemental Food Program for Women, Infants, and Children." *The American Journal of Clinical Nutrition*, vol. 32, March 1979.
Egbuonu, Lisa, and B. Starfield. "Inadequate Immunization and the Prevention of Communicable Diseases." *In The Effectiveness of Medical Care: Validating Clinical Wisdom*, edited by Barbara Starfield. Baltimore: Johns Hopkins University Press, 1985.

Eichorn, Dorothy H. "Head Start Profiles of Program Effects on Children: Motor Development." Westport, CT: Mediax Associates, 1978.

Eisen, M., C.A. Donald, J.E. Ware, and R.H. Brook. *Conceptualization and Measurement of Health for Children in the Health Insurance Study*. Santa Monica, CA: The Rand Corporation, 1980.

Fenson, Larry. "Parent Report: A Neglected Source of Data." Paper presented at the biennial meeting of the Society for Research in Child Development, New Orleans, March 1993.

Fenson, Larry, Phillip S. Dale, J. Steven Reznick, Donna Thal, Elizabeth Bates, Jeffrey P. Hartung, Steve Pethick, and Judy S. Reilly. *MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. San Diego: Singular Publishing Group, 1993.

Gresham, Frank M., and Stephen N. Elliott. *Social Skills Rating System: Manual*. Circle Pines, MN: American Guidance Service, 1990.

Hanushek, Eric A. "The Impact of Differential Expenditures on School Performance." *Educational Researcher*, vol. 18, no. 4, May 1989, pp. 45-51, 62.

Hauser-Cram, Penny, Donald E. Pierson, Deborah K. Walker, and Terrence Tivnan. *Early Education in the Public Schools: Lessons from a Comprehensive Birth-to-Kindergarten Program*. San Francisco: Jossey-Bass, 1991.

Helburn, Suzanne, et al. "Cost, Quality, and Child Results in Child Care Centers: Executive Summary." Denver: University of Colorado at Denver, 1995.

Hodges, W., A. Branden, R. Feldman, J. Follins, J. Love, R. Sheehan, J. Lumbley, J. Osborn, R.K. Rentfrow, J. Houston, and C. Lee. *Follow Through: Forces for Change in the Primary Schools*. Ypsilanti, MI: High/Scope Press, 1980.
Hofferth, S.L., A. Brayfield, S.G. Deich, and P. Holcomb. "The National Child Care Survey 1990." Washington, DC: Urban Institute, 1991.

Hollis[illegible], Robinson G., and Jennifer Hill. "Problems in the Evaluation of Community-Wide Initiatives." In *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, edited by James P. Connell, Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss. Washington, DC: The Aspen Institute, 1995.

Howes, Carollee, Deborah A. Phillips, and Marcy Whitebook. "Thresholds of Quality: Implications for the Social Development of Children in Center-Based Child Care." *Child Development*, vol. 63, 1992, pp. 449-460.

Infant Health and Development Program. "Enhancing the Results of Low-Birth-Weight, Premature Infants: A Multisite, Randomized Trial." *Journal of the American Medical Association*, vol. 263, no. 2, June 13, 1990, pp. 3035-3042.

Kagan, Sharon L., Evelyn Moore, and Sue Bredekamp (Eds.). "Reconsidering Children's Early Development and Learning: Toward Common Views and Vocabulary." Washington, DC: National Education Goals Panel, June 1995.

Koblinsky, S.A., M.L. Taylor, and L.W. Douglass. "Relationships Between Maternal, Child, and Shelter: Characteristics and Homeless Children's Developmental Status." College Park, MD: University of Maryland, 1995.

Kresh, Esther. "The Effects of Head Start: What Do We Know?" Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth and Families, 1993

Krauskopf, James A. "Overcoming Obstacles to Implementing Reform of Family and Children's Services." Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Chicago, October 1994.

Laosa, Luis M. "Evaluating the Effects of Head Start on Children's Cognitive Development." Westport, CT: Mediax Associates, 1978.

Larner, Mary. "Realistic Expectations: Review of Evaluation Findings." *In Fair Start for Children: Lessons Learned from Seven Demonstration Projects*, edited by Mary Larner, Robert Halpern, and Oscar Harkavy. New Haven: Yale University Press, 1992.

Lazar, I., R. Darlington, H. Murray, J. Royce, and A. Snipper. "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies." *Monographs of the Society for Research in Child Development*, vol. 47, nos. 2-3 (series no. 195), 1982.

Love, John M. "Indicators of Problem Behaviors and Problems in Early Childhood." Invited paper prepared for the Conference on Indicators of Children's Well-Being, National Institute of Child Health and Human Development, Bethesda, MD, November 17-18, 1994.

Love, John M., J. Lawrence Aber, and Jeanne Brooks-Gunn. "Strategies for Assessing Community Progress Toward Achieving the First National Education Goal." Princeton, NJ: Mathematica Policy research, Inc., October 1994.

Love, John M., Sally Wacker, and Judith Meece. "A Process Evaluation of Project Developmental Continuity, Interim Report II, Part B: Recommendations for Measuring Program Impact." Ypsilanti, MI: High/Scope Educational Research Foundation, June 1975.

Mallar, Charles D., and Rebecca A. Maynard. "The Effects of Income Maintenance on School Performance and Educational Attainment." In Research in Human Capital and Development: A Research Annual, vol. 2, edited by Ali Khan and Ismail Sirageldin. Greenwich, CT: JAI Press Inc, 1981.

Mathematica Policy Research, Inc. "Teenage Parent Demonstration 24-Month Follow-Up Questionnaire." Princeton, NJ: MPR, 1993.

Mathematica Policy Research, Inc. "Survey of Mothers and Children: Mother's Questionnaire for the Interactions and Developmental Processes Study." Princeton, NJ: MPR, 1991.

Maynard, Rebecca. "Some Aspects of Income Distribution: The Effects of the Rural Income Maintenance Experiment on the School Performance of Children." *American Economic Review*, vol. 67, no. 1, 1977, pp. 370-374.

Maynard, Rebecca, and Richard Murnane. "The Effects of a Negative Income Tax on School Performance: Results of an Experiment." *The Journal of Human Resources*, vol. 14, no. 4, 1979, pp. 463-476.

McCall, Robert B. "Head Start: Its Potential, Its Achievements, Its Future: A Briefing Paper for Policymakers." Pittsburgh, PA: University of Pittsburgh, Office of Child Development, 1993.

McCartney, Kathleen, Sandra Scarr, Deborah Phillips, Susan Grajek, and J. Conrad Schwarz. "Environmental Differences Among Day Care Centers and Their Effects on Children's Development." In *Day Care: Scientific and Social Policy Issues*, edited by Edward F. Zigler and Edmund W. Gordon. Boston: Auburn Press, 1982.

McKey, R.H., L. Condelli, H. Gamson, B. Barrett, C. McConkey, and M. Plantz. *The Impact of Head Start on Children, Families, and Communities: Final Report of the Head Start Evaluation, Synthesis, and Utilization Project.* Washington, DC: U.S. Department of Health and Human Services, June 1985.

Mediax Associates. "Accept My Profile: Perspectives for Head Start Profiles of Program Effects on Children." Westport, CT: Mediax Associates, 1980.

Meisels, Samuel J., Martha Stone Wiske, Laura W. Henderson, Dorothea B. Marsden, and Kimberly Browning. ESI 4-6: *Early Screening Inventory for Four Through Six Year Olds.* Revised, third edition. New York: Teachers College Press, 1988.

National Center for Education Statistics. *National Household Education Survey, 1993.* Washington, DC: NCES, 1993.

National Center for Health Statistics. "Design and Estimation for the National Health Interview Survey, 1985-94." *Vital and Health Statistics*, series 2, no. 110. Bethesda, MD: U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, August 1989.

Olds, David L., and Harriet Kitzman. "Can Home Visitation Improve the Health of Women and Children at Environmental Risk?" *Pediatrics*, vol. 86, 1990, pp. 108-116.

Peterson, J.L., and N. Zill. "Marital Disruption, Parent-Child Relationships, and Behavioral Problems in Children." *Journal of Marriage and the Family*, vol. 48, 1986, pp. 295-307.

Pierson, D.E., D.K. Walker, and T. Tivnan. "A School-Based Program from Infancy to Kindergarten for Children and Their Parents." *Personnel and Guidance Journal*, vol. 62, no. 8, 1984, pp. 448-455.

Raizen, Senta, and S.B. Bobrow. *Design for a National Evaluation of Social Competence in Head Start Children.* Santa Monica, CA: Rand Corporation, 1974.

Ramey, Craig T., and Frances A. Campbell. "Poverty, Early Childhood Education, and Academic Competence: The Abecedarian Experiment." In *Children in Poverty: Child Development and Public Policy*, edited by Aletha C. Huston. New York: Cambridge University Press, 1991.

Ramey, Craig T., Donna M. Bryant, and Tanya M. Suarez. "Preschool Compensatory Education and the Modifiability of Intelligence: A Critical Review." In *Current Topics in Human Intelligence*, edited by D. Detterman. Norwood, NJ: Ablex Publishing Corporation, 1983.

Raver, C. Cybele, and Edward F. Zigler. "Three Steps Forward, Two Steps Back: Head Start and the Measurement of Social Competence." *Young Children*, vol. 46, no. 4, 1991, pp. 3-8.

Schorr, Lisbeth B. "The Case for Shifting to Results-Based Accountability, with a Start-Up List of Outcome Measures with Annotations." Washington, DC: Center for the Study of Social Policy, Fall 1994.

Schwartz, Donald F., Jeanne Ann Grisso, Carolyn Miles, John H. Holmes, and Rudolph L. Sutton. "An Injury Prevention Program in an Urban African-American Community." *American Journal of Public Health*, May 1993, vol. 83, no. 5, pp. 675-680.

Schweinhart, Lawrence J., Helen V. Barnes, and David P. Weikart. "Significant Benefits: The High/Scope Perry Preschool Study Through Age 27." *Monographs of the High/Scope Educational Research Foundation*, no. 10. Ypsilanti, MI: High/Scope Educational Research Foundation, 1993.

Shepard, Lorrie, and M.L. Smith (Eds.). *Flunking Grades: Research and Policies on Retention.* New York: Palmer Press, 1989.

Starfield, Barbara. "The Effectiveness of Medical Care: Validating Clinical Wisdom." Baltimore: Johns Hopkins University Press, 1985, pp. 48-57.

St.Pierre, Robert, Jean Layzer, and Helen Barnes. "Variation in the Design, Cost, and Effectiveness of Two-Generation Programs." Paper prepared for the Eighth Rutgers Invitational Symposium on Education: New Directions for Policy and Research in Early Care and Education, Princeton, NJ, October 1994.

Thornton, Craig, John M. Love, and Alicia L. Meckstroth. "Community-Level Measures for Assessing the Status of Children and Families." Princeton, NJ: Mathematica Policy Research, December 1994.

Yoshikawa, Hirokazu. "Prevention as Cumulative Protection: Effects of Early Family Support and Education on Chronic Delinquency and Its Risks." Psychological Bulletin, vol. 115, 1994, pp. 28-54.

Zill, Nicholas. *Behavior Problems Index.* Washington, DC: Child Trends, Inc., 1990.

Zill, Nicholas, and Charlotte A. Schoenborn. "Developmental, Learning, and Emotional Problems: Health of Our Nation's Children, United States, 1988." *Advance Data from Vital and Health Statistics*, No. 190. Hyattsville, MD: National Center for Health Statistics, November 16, 1990.

TABLE 1

HEALTH OUTCOMES AND MEASURES: INCIDENCE OF PREVENTABLE DISEASES AND DISABILITIES IN CHILDREN

Intermediate-Term Outcomes Expected	Long-Term Outcomes Expected	Measures Recommended	Data Source	Evidence of Sensitivity to Community Interventions	Strengths [Liabilities] of the Measure	Policy Relevance
--	Decrease in postneonatal mortality rates or in race/ethnicity differential in postneonatal mortality rates	Postneonatal mortality rate for all children and by race/ethnicity	Birth certificate	Some evidence that postneonatal mortality is responsive to changes in the prevention of medical care (Starfield 1985)	[Linked birth and death records files are available from states with a long time lag.]	Reduced infant mortality
--	Decrease in preventable infant mortality	Postneonatal mortality rate > 2/1000	Birth certificate	Some evidence that postneonatal mortality is responsive to changes in the prevention of medical care (Starfield 1985)	[Linked birth and death records files are available from states with a long time lag.]	Reduced infant mortality
Reduced number of cases of diseases for which immunization is available--pertussis, polio, measles, mumps, or rubella	Reduced number of cases of diseases for which immunization is available	Number of children with diseases listed	Public health records and/or survey	Immunization found to decrease the incidence of vaccine-preventable communicable diseases (Egbuonu and Starfield 1985)	[Incidence of vaccine-preventable communicable diseases is low and outcomes are thus not appropriate for statistical analysis.]	Reduced health care costs, improved physical health
--	Increased use of safety precautions to reduce accidents and unintentional injury	Proportion of families using injury prevention: seat belts, ipecac syrup in house, working smoke alarms	Survey	An injury prevention program in a poor African American community resulted in increased proportion of homes with functioning smoke detectors, syrup of ipecac, safely stored medications, and reduced electrical and tripping hazards (Schwartz et al. 1993).		Reduced childhood morbidity and mortality

SOURCE: Thornton, Love, and Meckstroth (1994).

TABLE 2

HEALTH OUTCOMES AND MEASURES: OVERALL HEALTH OF CHILDREN

Intermediate-Term Outcomes Expected	Long-Term Outcomes Expected	Measures Recommended	Data Source	Evidence of Sensitivity to Community Interventions	Strengths [Liabilities] of the Measure	Policy Relevance
More positive parent perceptions of child's health status	Continued improvement in parent perceptions of child health status	NHIS question on mother's rating of child's health	Survey	Home visit programs improve maternal health and reduce risk of infant health problems (Olds and Kitzman 1990).	Scale widely used and useful in assessing current health status, changes in health status from treatment, and prediction of mortality	Improved overall health of children
Fewer children with functional limitations due to health conditions	Continued reduction in children with functional limitations due to health conditions	RAND Health Perceptions Scale	Survey	No effects detected of an early intervention program for low-birthweight, preterm infants (Infant Health and Development Program 1990).	Established scale with high internal consistency and good concurrent validity with other reported measures of health	Improved functional status of children
Fewer children with morbidities or serious morbidities	Fewer children with morbidities or serious morbidities	Morbidity Index, Serious Morbidity Index	Survey	Few effects detected of an early intervention program for low-birthweight, preterm infants (Infant Health and Development Program 1990).	Combination of individually rare events to give a more sensitive outcome measure	Improved physical health of children
	Increase in proportion of children who are within age-appropriate height and weight norms	Height and weight relative to age norms--percentiles and z-scores	School registration records	WIC participation by children reduced the percentage of children below the 10th percentile of height and weight for their age (Edozien et al. 1979).	Reliable and valid measures easily available from direct measurements or medical records review (Shankoff 1992).	Improved health of children

SOURCE: Thornton, Love, and Meckstroth (1994).

TABLE 3

CHILD DEVELOPMENT OUTCOMES AND MEASURES

Intermediate-Term Outcomes Expected	Long-Term Outcomes Expected	Measures Recommended (Age Range)	Data Source	Evidence of Sensitivity to Community Interventions	Strengths [Liabilities] of the Measure	Policy Relevance
Participation by Children in Early Education and Care Programs Before Kindergarten						
Increase in enrollments in early education and care programs	Continued increase in enrollments in early education and care programs	Enrollment records (ages 1-5 years)	Head Start records, child care system records School registration	Direct link to outreach activities	Obtainable from program databases [Data may not represent community as a whole.] [Some communities may have data in a number of different databases.]	Availability and use of child care are associated with parents' ability to work and reduction in welfare dependence.
		National Household Education Survey (NHES) items on types of programs attended, length of attendance, etc. (ages 3-6 years)	Survey		Used in 1993 NHES	Participation in early childhood programs expected to improve school readiness
Longer waiting lists for early childhood programs	Increased supply of early childhood programs; perhaps shorter waiting lists	Provider and service worker assessments of adequacy of supply (ages 1-5 years)	Program records			
Language Development						
Increased levels of receptive, expressive, and productive language	Continued increase in levels of receptive, expressive, and productive language	MacArthur Communicative Development Inventories--CDI--Short Form (ages 1-3 years)	Survey	Cognitive and language development are enhanced by such factors as involvement in quality child care and early education programs and increased levels of stimulation in the home (McCartney et al. 1982; Ramey et al. 1983).	Short form highly correlated with full CDI, which has strong construct validity--strong correlations between CDI vocabulary production scores and laboratory measures; mothers' reports of onset of grammatical functions closely parallel language development theory Age norms available [although may not be applicable to low-income samples] Used in NIH Infant Day Care Study [Norming sample did not have large numbers of minority group members.] [Limited to 8- to 36-month age range]	Language skills are central to school readiness and success in school. Relates to National Education Goal One: domain of language usage

BEST COPY AVAILABLE

TABLE 3 *(continued)*

Intermediate-Term Outcomes Expected	Long-Term Outcomes Expected	Measures Recommended (Age Range)	Data Source	Evidence of Sensitivity to Community Interventions	Strengths [Liabilities] of the Measure	Policy Relevance
School-Related Knowledge and Skills						
Increases in school-related knowledge and skills	Increases in school-related knowledge and skills	NHES items on knowledge of colors, letters, numbers, and writing (3-5)	Survey	School-related knowledge is enhanced by preschool program participation and improved family functioning (Barnett 1992).	Used in 1993 NHES	School readiness: relates to cognitive and general knowledge
		NHES items (5-6)	School registration			
		Early Screening Inventory (ESI) (4-6)	School registration		Good norms, strong reliability, has companion parent rating form	
Improved motor development and coordination	Improved motor development and coordination	NHES items on buttoning, holding a pencil, writing, and drawing; motor coordination (3-5)	Survey	Preschool programs can enhance school performance (Schweinhart et al. 1993).	Used in 1993 NHES	School readiness: relates to physical well-being and motor development
		ESI (4-6)	School registration		Good norms and reliability	
Social Well-Being						
Increased levels of cooperation, assertion, and responsibility, and increased degree of self-control	Increased levels of cooperation, assertion, and responsibility, and increased degree of self-control	Social Skills Rating System--Parent Form (SSRS) (3-5)	Survey	Quality child care improves social development (Howes et al. 1992).	Standardized on national sample of 5,000 children, racially, ethnically, and socioeconomically mixed	Social skills are important for successful adjustment to kindergarten; relates to social and emotional development and domain of approaches toward learning
		SSRS--Teacher Form (5-6)	School registration		Emphasizes positive behaviors	
Behavior Problems						
Lowered levels of headstrong, antisocial, anxious, depressed, and overly dependent behavior	Lowered levels of headstrong, antisocial, anxious, depressed, and overly dependent behavior	Behavior Problems Index--BPI (2-5)	Survey	Parent ratings on BPI distinguish children who have and have not received psychological help (Zill 1990); children in families with low levels of conflict experience fewer behavior problems (Peterson and Zill 1986); children with preschool experience have fewer classroom behavior problems (Pierson et al. 1984).	Designed as parent-report measure BPI is widely used (JOBS, NLSY-CS, NHIS) Strong psychometric properties	Behavior problems can interfere with school readiness and success Children with fewer behavior problems are less likely to require mental health services (Achenbach et al. 1991). Relates to social and emotional development

SOURCE: Thornton, Love, and Meckstroth (1994).

BEST COPY AVAILABLE

TABLE 4

SCHOOL PERFORMANCE OUTCOMES AND MEASURES

Intermediate-Term Outcomes Expected	Long-Term Outcomes Expected	Measures Recommended	Data Source	Evidence of Sensitivity to Community Interventions	Strengths [Liabilities] of the Measure	Policy Relevance
			Remediation			
--	Decreased proportions of children identified as developmentally delayed at kindergarten entry	Scores on developmental screening tests	Kindergarten registration forms	Enrollment in preschool programs and increased levels of intellectual stimulation in the home shown to decrease developmental delays (Ramey et al. 1983)	Inexpensive to collect (if available) [Subject to possible inconsistencies in school registration procedures]	Developmental delays are major source of increased educational costs.
--	Reductions in number of children assigned to special education programs	Number and proportion of children with special education assignments	School system records	Enrollment in preschool programs and increased levels of intellectual stimulation in the home shown to decrease special education placements (Barnett 1992; Schweinhart et al. 1993)	Inexpensive to collect (if available) [Subject to inconsistencies in school records]	Special education services are very costly.
			Attendance			
Improved attendance	Improved attendance	Days absent/present	School records	Improvements in the home environment can improve school attendance (Maynard 1977).	Inexpensive to collect [Subject to inconsistencies in school records]	Poor attendance interferes with school completion.
			Grade Progression			
--	Reductions in rates of children retained in grade	Proportion of students over age for grade	School records	Quality preschool experience reduces rate of grade retentions (Lazar et al. 1982; Schweinhart et al. 1993).	Inexpensive to collect [Subject to variation in school retention policies]	Grade retentions result in increased education costs.
Decreased school dropout rates	Continued decrease in school dropout rates	School dropout rate	School records	Improvements in the home environment can increase probability of completing high school and increasing educational attainment (Mallar and Maynard 1981). However, rigorously evaluated school-based programs for teenagers have shown little effect.		Dropping out of high school can decrease future earnings potential.

BEST COPY AVAILABLE

TABLE 4 *(continued)*

Intermediate-Term Outcomes Expected	Long-Term Outcomes Expected	Measures Recommended	Data Source	Evidence of Sensitivity to Community Interventions	Strengths [Liabilities] of the Measure	Policy Relevance
School Achievement						
Improved basic skills and academic achievement	Improved basic skills and academic achievement	Screening test scores	School system records	Cognitive and social performance improved by preschool attendance (see child development outcomes)	Inexpensive to collect (if available) [Subject to inconsistencies in school registration procedures]	Related to National Education Goals
–	–	Standardized achievement test scores	School system records	School achievement improved by attendance in quality preschool programs (Barnett 1992; Lazar et al. 1982), family background factors (Hanushek 1989), improvements in the home environment (Maynard 1977; Maynard and Murnane 1979; Mallar and Maynard 1981), infant and preschool home environments (Bradley and Caldwell 1978, 1984)		

SOURCE: Thornton, Love, and Meckstroth (1994).

BEST COPY AVAILABLE

TABLE 5

SCHOOL-ENTRY ASSESSMENT STRATEGIES FOR THE READINESS DIMENSIONS OF THE FIRST NATIONAL EDUCATION GOAL

Readiness Indicator	Measure
Access to High-Quality and Developmentally Appropriate Preschool Programs	
Increased enrollments in early care and education programs	Questions on types of programs attended (Head Start, nursery school, state prekindergarten, center day care, family day care), age of first attendance, and duration of attendance[a]
Improved quality of early care and education programs	Questions on program's daily and weekly schedule, group size, and child-staff ratio[a]
Increased stability of child care arrangements	Questions on number of different early care and education settings experienced[a,b]
Increased percentage of high-risk children enrolled in early intervention programs	Question on enrollment in early intervention programs[c]
Every Parent Will Be a Child's First Teacher and Devote Time Each Day to Helping His or Her Preschool Child Learn	
Increased amount of time spent with child in intellectually challenging activities	Questions on frequency of reading, storytelling, teaching activities, playing games, discussing science or nature, etc.[a]
Increased number of educational materials in the home	Questions on number of books, games, and other educational materials in the home[a]
Increased regulation of children's television viewing	Questions on regulation (rules covering content and hours) of children's television viewing[a]
Increased enriching experiences outside the home provided by parents	Questions on frequency of visits to libraries, museums, zoos, plays, concerts, churches, cultural organizations, and games or sports[a]
Parents Will Have Access to the Training and Support They Need	
Increased attendance at parenting classes in the community	Questions on attendance at classes[d,e]
Increased availability of parenting classes, social clubs, parent groups, counseling opportunities, social service agencies, and other supports	Questions on knowledge of and access to parenting classes, social clubs, parent groups, counseling opportunities, social service agencies, and other supports[d,e]
Children Will Receive the Nutrition and Health Care Needed to Arrive at School with Healthy Minds and Bodies	
Reduced percentage of low-birthweight babies	Questions on child's weight at birth[f]
Increased access to prenatal care	Question on number of prenatal care visits[f]
Increased percentage of children receiving regular medical care	Questions on regular source of routine care[f]
Increased percentage of children receiving regular well-child examinations	Questions on routine health checkup in past 12 months[f]
Decreased use of emergency room for nonemergency care	Question on emergency room use[f]
Increased percentage of children receiving immunizations at appropriate ages	Questions on completed immunizations[f]
Increased percentage of children having private or public health insurance	Questions on health insurance[f]
Increased percentage of children having regular vision and hearing screenings	Questions on vision and hearing screenings in past 12 months[f]
Decreased percentage of children having previously undetected vision and hearing problems	Questions on children referred for treatment[f]

TABLE 5 (continued)

Readiness Indicator	Measure
Increased percentage of children having completed dental checkup	Questions on dental visit in past 12 months[f]
Increased percentage of children having nutrition screening before kindergarten entry	Questions on nutrition screening[f]
Increased percentage of children referred before kindergarten for treatment of mild asthma, tuberculosis, cerebral palsy, mental retardation, autism, or other pervasive developmental disorders	Questions on treatment referrals[f]
Physical Well-Being and Motor Development	
Physical Well-Being	
Increased overall health status of children	Child's health rating[a,f]
Decreased percentage of children having functional limitations because of health conditions	Ratings of child's functional limitations[g]
Reduction in percentage of children with morbidities or serious morbidities	Child's health rating[h]
Decreased number of hospitalizations	Questions on hospitalizations[f]
Increased percentage of children within age-appropriate height and weight norms	Items on height and weight relative to age norms (by direct examination or medical record review)
Motor Development	
Improved fine-motor development and coordination	Items on block-building, draw-a-person, and copying forms[i]
Improved gross-motor skills	Items on gross-motor/body-awareness scale[i]
Social and Emotional Development	
Social Development	
Increased levels of assertion	Scores on assertion scale[j]
Decreased levels of aggressive behavior, dependence, and headstrong behavior	Scores on scales measuring aggressive behavior, dependence, and headstrong behavior[k]
Increased cooperation and ability to help, communicate problems, and follow rules	Scores on cooperation scale[j]
Emotional Development	
Reduced levels of anxiety and depression	Scores on anxiety and depression scales[k]
Approaches Toward Learning	
Self-Control and Self-Regulation	
Increased ability to control temper, attend to instructions, and take turns	Scores on self-control scale[j]
Task Attention	
Increased ability to attend to task and to remember auditory material received	Items on auditory sequential memory scale[i]
Language Usage	
Verbal Expression	
Increased expressive language, speaking ability, and ability to describe objects	Items on verbal expression scale[i]

BEST COPY AVAILABLE

TABLE 5 (continued)

Readiness Indicator	Measure
Cognition and General Knowledge	
Visual Sequential Memory	
Increased ability to remember what is seen	Items on visual sequential memory[i]
Number Concepts	
Increased ability to count and to understand simple quantitative concepts	Items on number concept[i]
Verbal Reasoning	
Increased relational-thinking ability	Items on verbal reasoning[i]
General Knowledge	
Increased school-related knowledge and skills	Questions on knowledge of colors, letters, numbers, and writing[a]
Family Background, Demographics, and Contextual Variables	
Mother's education Father's education Mother's age at birth of first child Household structure Household income Employment status of mother and father Number of residential moves Length of residence in community Child's age and gender Child's disabilities Race/ethnicity of child Child's contacts with father (if father not in home) Number and ages of siblings Language(s) spoken in the home Neighborhood characteristics School characteristics	Questions selected from various instruments[e]

SOURCE: Love, Aber, and Brooks-Gunn (1994).

[a]National Household Education Survey (NHES:93) (NCES 1993).

[b]National Child Care Survey 1990 (Hofferth et al. 1991).

[c]Interactional and Developmental Processes study questionnaire (MPR 1991).

[d]Teenage Parent Demonstration 24-Month Follow-Up questionnaire (MPR 1993).

[e]Measures to be selected.

[f]National Health Interview Survey (NHIS) Child Health Supplement (NCHS 1989).

[g]Rand Health Perceptions Scale (Eisen et al. 1980).

[h]Morbidity Index and Serious Morbidity Index (Brooks-Gunn et al. 1994).

[i]Early Screening Inventory (ESI) (Meisels et al. 1988).

[j]Social Skills Rating System (SSRS) (Gresham and Elliott 1990).

[k]Behavior Problems Index (BPI) (Zill 1990).

TABLE 6

CHARACTERISTICS OF FOUR ILLUSTRATIVE CHILD OUTCOME MEASURES

Feature/Characteristic	Behavior Problems Index (Zill 1990)	Early Screening Inventory (Meisels et al. 1988)	MacArthur Communicative Development Inventories (Fenson et al. 1993)	Social Skills Rating System--Elementary Teacher Form (Gresham and Elliott 1990)
Outcomes measured	Behavior problems labeled as • Headstrong • Aggressive • Anxious • Depressed • Immature/dependent	Visual-motor/adaptive • Draw-a person • Fine-motor control • Eye-hand coordination • Visual-sequential memory • Reproducing 2- and 3-dimensional visual structures Language and cognition • Comprehension • Verbal expression • Reasoning • Counting • Remembering auditory sequences Gross-motor/body awareness • Large-muscle coordination • Balancing, hopping, skipping • Imitating body positions from visual cues	Words and Gestures (infants)[a] • Phrases • Vocabulary comprehension • Vocabulary production • Early gestures • Later gestures Words and Sentences (toddlers) • Vocabulary production • Irregular nouns and verbs • Sentence complexity	Cooperation Assertion Self-control Externalizing behaviors Internalizing behaviors Hyperactivity Academic competence
Age range assessed	2-5 years	4-6 years	Infant form: 8-16 months Toddler form: 16-30 months	5-11 years (grades K-6)
Type of assessment	Rating by parent	Structured individual assessment by teacher	Parent report	Rating by teacher
How administered	Telephone interview, self-report, or in-person interview	Administered by teacher	Self-report by parent	Written response

BEST COPY AVAILABLE

TABLE 6 *(continued)*

Feature/Characteristic	Behavior Problems Index (Zill 1990)	Early Screening Inventory (Meisels et al. 1988)	MacArthur Communicative Development Inventories (Fenson et al. 1993)	Social Skills Rating System--Elementary Teacher Form (Gresham and Elliott 1990)
Administration time	10 minutes	15-20 minutes	20-30 minutes	10 minutes
Standardization and norms	Data available on 17,110 children 17 years of age and under from the 1988 National Health Interview Survey of Child Health.	Normed on national sample of 2,746 children, 44 percent nonwhite.	Norming sample of 671 infants and 1,142 toddlers in New Haven, Seattle, and San Diego	Norming sample of 5,000 children--geographically, racially, and socioeconomically heterogenous (956 for the elementary teacher form); included 19 percent handicapped
Previous and/or current use	National Longitudinal Survey of Youth Child Supplements (1986, 1988) National Health Interview Survey: Child Health Supplement (1981, 1988) NICHD infant day care study JOBS child impact study	Evaluation of developmental status of homeless children (Koblinsky, Taylor, and Douglas 1995)	Currently used in NICHD infant day care study	Head Start-Public School Transition Demonstration evaluation Under consideration for use in the Early Childhood Longitudinal Study (NCES)
Strengths	Extensive national data available for comparative use Short and easy to administer	Strong predictive validity with the McCarthy Scales of Children's Abilities (Meisels et al. 1993) High interscorer and test-retest reliabilities Refers a high proportion of children actually at-risk, and excludes most of the children not at-risk (Meisels et al. 1993)	Good predictive validity (using observational data as criterion measure) Moderate to high internal consistency, depending on scale Good cross-form, cross-age stability during period of early and more rapid vocabulary expansion	Good psychometrics Norms for handicapped and nonhandicapped children Assesses positive social skills as well as problem behaviors
Liabilities	Programs may not like the negativity of focus on problem behaviors	Administration time could constitute significant burden if a teacher is asked to assess multiple children	Only 13 percent of the norming sample were minority group members; 78 percent of parents had some college education or a college degree.	Administration time could constitute significant burden if a teacher is asked to rate multiple children

TABLE 6 *(continued)*

Feature/Characteristic	Behavior Problems Index (Zill 1990)	Early Screening Inventory (Meisels et al. 1988)	MacArthur Communicative Development Inventories (Fenson et al. 1993)	Social Skills Rating System--Elementary Teacher Form (Gresham and Elliott 1990)
Comments	Ratings by teacher or other knowledgeable adult also possible		Administration by interview also possible. Short form is being field tested. May be suitable for older children who are developmentally delayed.	

[a]This list represents the subscales of the CDI. Various variables can be generated as outcome measures, e.g., age at which x percent of the sample uses plurals, possessives, progressive, and past tense.

BEST COPY AVAILABLE